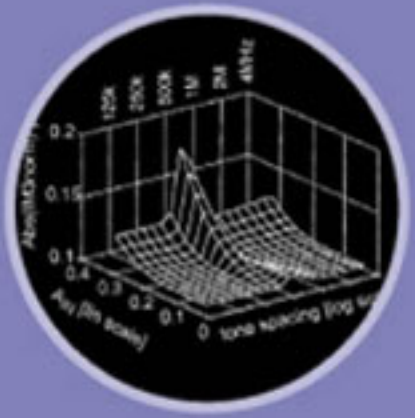
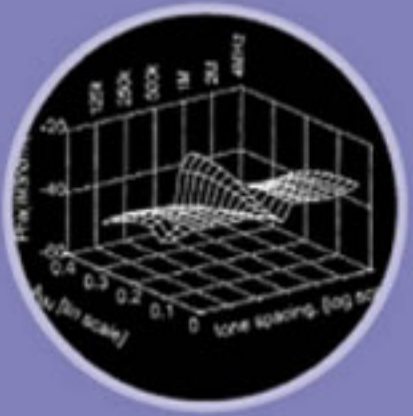
DISTORTION IN RF POWER AMPLIFIERS

Joel Vuolevi

Timo Rahkonen





Distortion in RF Power Amplifiers

For a listing of recent titles in the *Artech House Microwave Library,* turn to the back of this book.

Distortion in RF Power Amplifiers

Joel Vuolevi
Timo Rahkonen

Artech House
Boston • London
www.artechhouse.com

Library of Congress Cataloging-in-Publication Data
Vuolevi, Joel.
Distortion in RF power amplifiers / Joel Vuolevi, Timo Rahkonen.
p. cm. — (Artech House microwave library)
Includes bibliographical references and index.
ISBN 1-58053-539-9 (alk. paper)
1. Power amplifiers. 2. Amplifiers, Radio frequency. 3. Electric distortion—Prevention.
I. Rahkonen, Timo. II. Title. III. Series.

TK7871.58.P6V79 2003
621.384'12—dc21

2002043669

British Library Cataloguing in Publication Data
Vuolevi, Joel
Distortion in RF power amplifiers. — (Artech House
microwave library)
1. Power amplifiers 2. Amplifiers, Radio frequency 3. Radio—
Interference
I. Title II. Rahkonen, Timo
621.3'8412

ISBN 1-58053-539-9

Cover design by Gary Ragaglia

685 Canton Street
Norwood, MA 02062

International Standard Book Number: 1-58053-539-9
Library of Congress Catalog Card Number: 2002043669

10 9 8 7 6 5 4 3 2 1

Contents

Acknowledgments

Many persons and organizations deserve warm thanks for making this book a reality. To mention a few, Jani Manninen has made many of the measurements and test setups presented in this book, Janne Aikio contributed much to the characterization measurement techniques, and Antti Heiskanen contributed to the higher order Volterra analysis. Mike Faulkner and Lars Sundström originally introduced us to this linearization business. Veikko Porra and Jens Vidkjaer pointed out several important topics to probe further. The grammar and style of this book and the original publications on which it is mostly based have been checked by Janne Rissanen, Malcolm Hicks, and Rauno Varonen. Also, David Choi spent a lot of time with the text to make it more readable and fluent.

The financial and technical support of TEKES (National Technology Agency of Finland), Nokia Networks, Nokia Mobile Phones, Elektrobit Ltd, and Esju Ltd is gratefully acknowledged. The work has also been supported by the Graduate School in Electronics, Telecommunications and Automation (GETA) and the following foundations: Nokia Foundation, Tauno Tönningin säätiö, and Tekniikan edistämissäätiö.

Last but most important, we would like to thank our very nearest: Katja, Aleksi, Kaarina, and Antti Vuolevi, Paula Pesonen, and Kaija, Heikki, and Ismo Rahkonen.

Chapter 1

Introduction

1.1 Motivation

This book is about nonlinear distortion in radio frequency (RF) power amplifiers (PAs). The purpose of the PA is to boost the radio signal to a sufficient power level for transmission through the air interface from the transmitter to the receiver. This may sound simple, but it involves solving several contradicting requirements, the most important of which are linearity and efficiency. Unfortunately, these requirements tend to be mutually exclusive, so that any improvement in linearity is usually achieved at the expense of efficiency, and vice versa.

To avoid interfering with other transmissions, the transmission must stay within its own radio channel. If the modulated carrier has amplitude variations, any nonlinearity in the amplifier causes spreading of the transmitted spectrum (so-called spectral regrowth). This effect can be reduced by using constant-envelope modulation techniques that unfortunately have quite low data rate/bandwidth ratio. When using more efficient digital modulation techniques, the only solution is to design the amplifiers linear enough.

The efficiency is defined as a ratio of the generated RF power and the drawn dc power. In modern radio telecommunication systems, the design of linear and efficient radio frequency power amplifier presents one of the most challenging design problems. In general, relatively high transmit power levels are needed, and the power consumption of the PA easily dominates over all other electronics and digital processing in a mobile terminal. Therefore, high efficiency is essential to extend the operation time of the terminals. In fixed-point wireless nodes (e.g., in base stations), efficiency is also important, because the transmitted power levels are essentially higher than in terminals.

1.2 Historical Perspective

In first-generation systems, such as the Nordic Mobile Telephone (NMT) or Advanced Mobile Phone Service (AMPS), the RF signal was frequency modulated (FM). Highly efficient PAs are possible in FM systems because of the fact that no information is encoded in the amplitude component of the signal. Even so, the PA of a mobile phone consumed as much as 85% of the total system power at the maximum power level, thus limiting the on-time of the terminal.

Unlike wired line communications, wireless systems must share a common transmission medium. The available spectrum is therefore limited, and so channel capacity (i.e., the amount of information that can be carried per unit bandwidth) is directly associated with profit. The demand for greater spectral efficiency was addressed by the development of second-generation systems, where digital transmission and time domain multiple access (TDMA) is used, where multiple users are time multiplexed on the same channel. For example, in the Global System for Mobile Communications (GSM), eight calls alternate on the same frequency channel, resulting in cost-effective base stations. The GSM modulation scheme retains constant envelope RF signals, but the need for smooth power ramp up and ramp down of the allocated time-slot transmissions imposes some moderate linearity requirements. This reduces the efficiency of the amplifier, but it is compensated by the fact that the PA in the mobile node is only active one-eighth of the time. This, together with the smart idling modes, allows GSM handsets to achieve very long operating times.

The data transmission capacity of GSM is rather modest, so the obvious solution to increase the achievable bit rate was, as implemented in GSM-EDGE, to use several time slots for a single transmission and to replace the Gaussian minimum shift key (GMSK) modulation scheme with a spectrally more efficient 8-PSK that unfortunately has a varying envelope. So as wireless communication systems migrate towards higher channel capacity, more linear and, consequently, less efficient PAs have become the norm.

Finally, the third generation wideband code-division multiple access (WCDMA) packs tens of calls on the same radio channel simultaneously, differentiated only by their unique, quasi-orthogonal spreading codes. This allows flexible allocation of data rates, while tolerance to fading is improved by increasing the signal bandwidth to nearly 4 MHz. The advantages offered by the WCDMA, however, come at the expense of more stringent requirements for the PA. The code-multiplexed transmission occupies a much larger bandwidth than in the previous systems, while exhibiting tremendous variations in amplitude. Furthermore, in WCDMA,

the mobile transmits on a continuous time basis. Designing an economical PA for these requirements is an enormous engineering challenge.

The situation is not easier in the base stations, either, where the linearity requirements are tighter than in handsets. The trend is towards multicarrier transmitters where a single amplifier handles several carriers simultaneously, in which case the bandwidth, power level, and the peak power to average power ratio (crest factor) all increase. The efficiency of these kinds of power amplifiers is very low, and due to higher total transmitted power, this results in very high power dissipation and serious cooling problems.

1.3 Linearization and Memory Effects

The goal of this book is to improve the conceptual understanding needed in the development of PAs that offer sufficient linearity for wideband, spectrally efficient systems while still maintaining reasonably high efficiency. As already noted, efficiency and linearity are mutually exclusive specifications in traditional power amplifier design. Therefore, if the goal is to achieve good linearity with reasonable efficiency, some type of linearization technique has to be employed. The main goal of linearization is to apply external linearization to a reasonably efficient but nonlinear PA so that the combination of the linearizer and PA satisfy the linearity specification. In principle, this may seem simple enough, but several higher order effects seriously limit its effectiveness, in practice.

Several linearization techniques exist, and they are reviewed in Chapter 3; a much more detailed discussion can be found from [1-3]. Stated briefly, linearization can be thought of as a cancellation of distortion components, and especially as a cancellation of third-order intermodulation (IM3) distortion, and where the achieved performance is proportional to the accuracy of the canceling signals. Unfortunately, the IM3 components generated by the power amplifier are not constant but vary as a function of many input conditions, such as amplitude and signal bandwidth. Here, these bandwidth-dependent phenomena are called *memory effects*.

Smooth, well-behaved memory effects are usually not detrimental to the linearity of the PA itself. If the phase of an IM3 component rotates 10º to 20º, or if its amplitude changes 0.5 dB with increased tone spacing in a two-tone test, it usually does not have a dramatic effect on the adjacent channel power ratio (ACPR, i.e., the power leaking to the neighboring channel) performance of a standalone amplifier, nor is it especially of concern if the lower ACPR is slightly different from the upper one. However, the situation may be quite different if certain linearization

techniques are used to cancel out the intermodulation sidebands; in fact, the reported performance of some simple techniques may actually be limited not by the linearization technique itself, but by the properties of the amplifier – and especially by memory effects.

Different linearization techniques have different sensitivities to memory effects. Feedback and feedforward systems (see Section 3.2.2) are less sensitive to memory effects because they measure the actual output distortion, including the memory effects. However, predictive systems like predistortion and envelope elimination and restoration (EER) are vulnerable to any changes in the behavior of the amplifier, and memory effects may cause severe degradation in the performance of the linearizer.

However, there is no fundamental reason why predictive linearization techniques should be poorer than feedback or feedforward systems since the behavior of spectral components, though quite difficult to predict under varying signal conditions, is certainly deterministic. Thus, in theory, real time adaptation or feedback/feedforward loops are not strictly necessary, provided that the behavior of distortion components is known or can be controlled. The primary motivation of this book is to develop a power amplifier design methodology which yields PA designs that are more easily linearized. The approach taken here proposes that, by negating the relevant memory effects, the performance of simple linearization techniques that otherwise do not give sufficient linearization performance, can be significantly improved.

To achieve a significant linearity improvement by means of simple and low power linearization techniques requires detailed understanding of the behavior and origins of the relevant distortion components. This is a key theme that is carried on throughout this book. The actual linearization techniques themselves will not be discussed in detail, but instead, the fundamental aim of this book is to give the designer the crucial insights required to understand the origins of memory effects, as well as the tools to keep memory effects under control.

1.4 Main Contents of the Book

Obtaining meaningful data of signal bandwidth-dependent effects has been nearly impossible, as most commercially available RF power devices are supplied without simulation models, while those that are often fail even to fully reproduce the devices' I-V and Q-V curves. Hence, the predicted distortion characteristics from computer simulations is generally regarded as unsatisfactory; the results may be accurate within 5 dB, but this is not

sufficient for analyzing canceling linearization systems, where subdecibel accuracy is a prerequisite.

In laboratory measurements, the commonly used single-tone amplitude and phase distortion (AM-AM and AM-PM) characterization techniques actually have a zero bandwidth, and so they completely fail to capture bandwidth-dependent phenomena. Therefore, the accuracy of IM3 values resulting from AM-AM and AM-PM models suffers when attempting to model an amplifier that has memory effects. In addition, the AM-AM measurements also suffer from self-heating: The AM-AM measurements are performed using continuous wave (CW) signals, resulting in transistor junction temperatures quite different from those generated in practice, where modulated signals are applied to the PA.

This book presents several techniques that help understand, simulate, measure, and cancel memory effects. The subsequent chapters will provide a detailed discussion of the following topics:

1. A comparison between data available from AM-AM and AM-PM versus IM measurements. Normal single-tone AM-AM measurement has zero bandwidth, but it can be performed using a two-tone signal with variable tone spacing, as well. In this case, the same information about the nonlinearity of the device should be available in both the fundamental and IM3 tones, but the discussion will show that the large fundamental signal masks a considerable amount of fine variations in distortion in AM-AM measurements.
2. To study the phase variations of the IM3 tones, a three-tone measurement system will be presented.
3. Device modeling. Input-output behavioral models can be generated on the basis of a completed amplifier, but these do not yield any information to aid in design optimization. Instead, the analysis presented in this book models the transistor by replacing every nonlinear circuit element (input capacitance, g_m, and so forth) by the parallel combination of a linear circuit element (small-signal capacitance, small-signal g_m, and so forth) and a nonlinear current source. This leads to two important findings:
   a. There are several sources of distortion, and the distortion generated in any of these sources can undergo subsequent mixing processes, resulting in higher order distortion components than the degree of the nonlinearity suggests.

   b. Distortion is originally generated in form of current, which is converted to a voltage by terminal impedances. Thus, the phase and amplitude of the distortion components can be strongly influenced by the terminal impedances, and especially by the impedances of the biasing networks.

4. Based on the reasoning above, this book includes a review of a distortion analysis technique called Volterra analysis, which is based on placing polynomial distortion sources in parallel with linear circuit elements. The main benefits of this technique are:

   a. The dominant sources of distortion can be pinpointed;
   b. Phase relationships between distortion contributions can be easily visualized;
   c. A polynomial model can be accurately fitted to the measured data;
   d. The polynomial models can also be used in harmonic balance simulators.

5. This book also introduces some circuit techniques for reducing memory effects in power amplifiers. The standard method of minimizing memory effects involves attempting to maintain impedances at a constant level over all frequency bands. Unfortunately, other design requirements often interfere with this aim and cause memory effects. To address this problem, an active impedance synthesis technique is introduced, which can be used to drive impedances to their optimum values. What is more, this technique can be used for electrical and thermal memory effects.

6. Finally, the book presents a characterization technique for polynomial nonlinearities. Since many existing power transistor models are not sufficiently accurate in terms of distortion simulations, characterization measurements are the only way of obtaining this information. This is accomplished using pulsed *S*-parameter measurements over a range of terminal voltages and temperatures.

1.5 Outline of the Book

The main emphasis of this book is on developing a detailed understanding of the physics underlying distortion mechanisms, while keeping the mathematical formulations in a tractable form. To lay the groundwork for the analysis of nonlinear effects in RF power amplifiers, Chapter 2 discusses certain theoretical aspects related to amplifier circuits. Since RF power amplifiers are nonlinear, bandwidth-dependent circuits with

memory, it is important to define nonlinearity, bandwidth dependency, and memory, and to examine their associated effects. Chapter 2 also introduces a direct calculation method for deriving equations for the spectral components generated in such circuits. Due to its analytical nature, this method, based on the Volterra series, provides detailed information about distortion mechanisms in nonlinear systems. Later chapters of this book will describe the use of the method.

Chapter 3 first discusses memory effects from the linearization point of view. Some of the most common linearization techniques are presented, and then the chapter highlights the harmful memory effects in more detail, with a particular focus on electrical and thermal memory effects. Electrical memory effects are those caused by varying node impedances within a frequency band, while thermal memory effects are caused by dynamic variations in chip temperature. Both kinds of memory effects are analyzed by comparing a memoryless polynomial model with measurements of real power amplifier devices. Memory effects tend to be considered merely in terms of modulation frequency, but Chapter 3 also introduces mechanisms that produce memory effects as a function of signal amplitude. These mechanisms are referred to as amplitude domain memory effects.

Chapter 4 discusses transistor/amplifier models and introduces problems related to PA modeling. The amplifier models are classified as either behavioral or device-level models, which are based on some predefined, physically based functions or simply on empirical fitting functions. The Volterra model is an empirical model that is capable of providing component-level information that can be used for design optimization. The chapter also gives a derivation of the Volterra models for a common-emitter (CE) bipolar junction transistor (BJT) amplifier and a common-source (CS) metal-semiconductor field effect transistor (MESFET) amplifier. The models take into account the effects of modulation frequency, and temperature, and are therefore able to model memory effects. Moreover, IM products are presented as vector sums of each degree of nonlinearity, thereby providing insight into the composition of distortion, which is instrumental in design optimization.

Chapter 5 discusses the characterization of the Volterra model. The dc characterization is briefly discussed for the sake of clarity, before shifting the focus on a new technique based on a set of small-signal *S*-parameters measured over a range of bias voltages and temperatures.

Chapter 6 presents a new simulation technique that offers insight into both amplitude and modulation frequency-dependent memory effects. A new measurement technique is introduced that allows both the amplitude and the phase of the IM3 components to be measured, which is an

important improvement over measurements based merely on the fundamental signal or amplitude.

Chapter 7 introduces three techniques for canceling memory effects: impedance optimization, envelope filtering, and envelope injection. In addition, the chapter presents the source pull test setup for investigating the effects of out-of-band impedances. Then, a comparison is presented between envelope filtering and envelope injection techniques, and the superior compensation properties of the envelope injection technique are demonstrated. Finally, a detailed presentation of the envelope injection technique is given, and it is shown how both modulation frequency and amplitude domain effects can be compensated. A primary advantage of the memory effect cancellation approach is that the performance of a polynomial predistorter or other simple linearization technique can be significantly increased without a substantial increase in dc power consumption. Hence, good cancellation performance can be achieved by linearization techniques that consume little power, enabling the design of linear yet power-efficient PAs.

Finally, additional supporting information is collected in the appendixes. Appendixes A and B discuss the background and limits of the Volterra analysis. Appendix C includes a full list of transfer functions, describing the path from all of the distortion sources to a given node voltage in a common-emitter type single-transistor amplifier. Appendix D includes a brief description of some practical aspects of the measurement setups and the RF predistorter linearizer used in the measurements presented in Chapter 7.

References

[1] Raab, F., et al., "Power amplifiers and transmitters for RF and microwave," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 50, No. 3, 2002, pp. 814-826.

[2] Kenington, P. B., *High Linearity RF Amplifier Design,* Norwood, MA: Artech House, 2000.

[3] Cripps, S., *Advanced Techniques in RF Power Amplifier Design*, Norwood, MA: Artech House, 2002.

Chapter 2

Some Circuit Theory and Terminology

This chapter reviews the theoretical background needed for understanding nonlinear effects in RF power amplifiers. It begins comfortably by defining memory and linearity, and briefly reviewing phasor analysis and the most common ways to measure and define the amount of nonlinearity. It is also noted that nonlinear effects are more clearly and accurately seen as the structure of IM tones than as small AM-AM and AM-PM variations on top of the large fundamental signal. Sections 2.2 and 2.3 motivate the use of polynomial models, as the calculation of discrete tone spectrums in polynomial nonlinearities is easily done by convolving the original two-sided spectrums.

Section 2.4 defines the memory effects as in-band variation of the distortion: the behavior of intermodulation distortion at the center of the channel is different from that at the edge of the channel. Nonlinear analysis methods are very briefly discussed in Section 2.5, and the rest of the chapter concentrates on presenting Volterra analysis using what is known as the direct method or nonlinear current method. The method is very similar to linear noise analysis: Distortion is modeled as excess signal sources parallel to linear components. The main advantages of the Volterra analysis are that we get per-component information about the structure of distortion as well as the phase of these components, so that we can clearly see which distortion mechanisms are canceling each other and how to change the impedances to improve the cancellation, for example.

Finally, a simple example circuit is studied to see the analysis procedure, and the circuit-level presentation is briefly compared with a behavioral input-output model typically used in system simulations. The intention is to show that AM-PM can be modeled by an input-output polynomial with complex coefficients (or any complex function), but if the coefficients are fixed, it cannot predict bandwidth-dependent phenomena.

2.1 Classification of Electrical Systems

Electrical systems can be classified into four main categories as listed in Table 2.1: linear and nonlinear systems with or without memory. An example of a linear memoryless system is a network consisting of linear resistors. Addition of an energy storage element such as a linear capacitance causes memory, as a result of which a linear system with memory is introduced.

Nonlinear effects in electrical systems are caused by one or more nonlinear elements. A system comprising linear and nonlinear resistors is known as a memoryless nonlinear system. Nonlinear systems with memory, on the other hand, include at least one nonlinear element and one memory introducing element (or a single element introducing both).

Table 2.1
Classification of Electrical Systems

	Memoryless	With Memory
Linear	Linear resistance	Linear capacitance
Nonlinear	Nonlinear resistance	Nonlinear capacitance or nonlinear resistance and linear capacitance

2.1.1 Linear Systems and Memory

Any energy-storing element like a capacitor or a mass with thermal or potential energy causes memory to the system. This is seen from the voltage equation of a linear capacitance, for example:

$$v_C(t) = \frac{1}{C} \cdot \int_{-\infty}^{t} i(t') \cdot dt' \tag{2.1}$$

Here, the voltage at time t is proportional to all prior current values, not just to the instantaneous value. This is the reason why capacitances and inductances are regarded as memory-introducing circuit elements.

The well-known consequence of memory is that the time responses of the circuit are not instantaneous anymore, but will be convolved by the

impulse response of the system; in a system with long memory, the responses will be spread over a long period of time. This is illustrated in Figure 2.1(b) where the time domain output of a linear system of Figure 2.1(a) with and without memory is shown. Let the input signal be a ramp that settles to the normalized value of one. In a linear memoryless system, the output waveform is an exact, albeit attenuated (or amplified), copy of the input signal. If the system exhibits memory, the output waveform will be modified by the energy-storing elements.

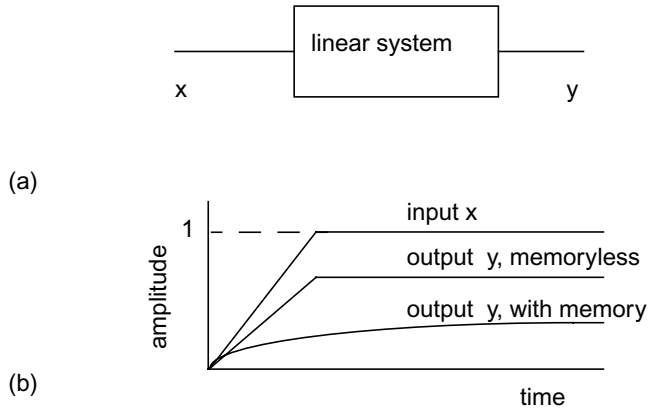


Figure 2.1 (a) Linear system and (b) its output in a time domain with and without memory.

In the frequency domain, the consequence of memory is seen as a frequency-dependent gain and phase shift of the signal. To analyze frequency-dependent effects, phasor analysis is commonly used: sinusoidal signals are written according to Euler's equation as a sum of two complex exponentials (phasors)

$$x = A_1 \cos(\omega_1 t + \phi_1) = \frac{A_1 e^{j\phi_1}}{2} \cdot e^{j\omega_1 t} + \frac{A_1 e^{-j\phi_1}}{2} \cdot e^{-j\omega_1 t} \quad , \quad (2.2)$$

where the time-dependent part models the rotating phase that can be frozen to a certain point in time (like t=0), and the complex-valued constant part contains both the amplitude A_1 and phase ϕ_1 information that fully describe a sinusoid with fixed frequency ω_1. The reader should note that in linear systems no new frequencies are generated, and the system is usually analyzed using positive frequency $+\omega_1$ only. In nonlinear analysis, new frequency components are generated, and both positive and negative phasors are needed to be able to calculate all of them, as we will see. Also, the fact that the complex phasors contain the phase information will turn out to be very handy when the cancellation of different distortion components is calculated.

The main advantage of phasor analysis (or using sinusoidal signals only, the derivatives and integrals of which are also sinusoids) is that the integrals and differentials involved in energy-storing elements reduce to multiplications or divisions with $j\omega$, where the imaginary number j means in practice a phase shift of +90º. This way differential equations are reduced to algebraic equations again, and normal matrix algebra is used to quickly solve the circuit equations. Table 2.2 reviews the device equations for basic components to be used in phasor analysis.

Table 2.2
Impedances and Admittances of Basic Circuit Elements

	Impedance Z = V/I	Admittance Y = I/V
L	$j\omega L$	$1 / (j\omega L) = -j / (\omega L)$
C	$1 / (j\omega C) = -j / (\omega C)$	$j\omega C$
R	R	$1 / R$

We see that energy-storing elements cause phase shift, while memoryless resistive circuits do not. This is further illustrated in Figure 2.2 where the impedance Z of a series RC network is shown in a complex plane as a vector sum of the impedances of $Z_R=R$ and $Z_C=1/j\omega C$, calculated at a certain value of ω. As Z_C is frequency-dependent, the magnitude and the phase of total impedance $R+1/j\omega C$ vary with frequency ω, which does not happen in a memoryless circuit.

Here, the total impedance of a series circuit was drawn as a vector sum of two contributions. Later we will construct the phasors of distortion tones as similar vector sums of different contributions.

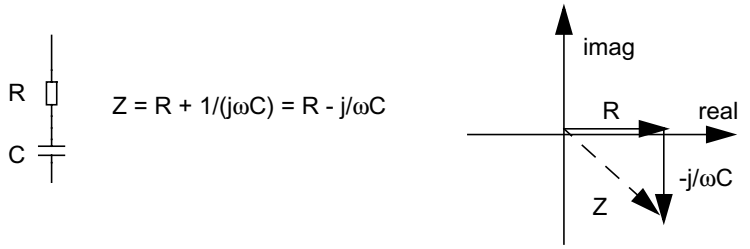


Figure 2.2 Impedance Z of a series connection of R and C shown as a vector sum of Z_R and Z_C.

2.1.2 Nonlinear Systems

Next, we discuss the nonlinear effects. A system is considered linear if the output quantity is linearly proportional to the input quantity, as shown by the dashed line in Figure 2.3. The ratio between the output and the input is called the gain of the system, and in accordance with the definition presented above, it is not affected by the applied signal amplitude. A nonlinear system, in contrast, is a system in which the output is a nonlinear function of the input (solid line) (i.e., the gain of the system depends on the value of the input signal). If the output quantity is a current, and the input quantity a voltage, Figure 2.3 represents a nonlinear conductance. If the output quantity is changed to a charge, nonlinear capacitance is presented.

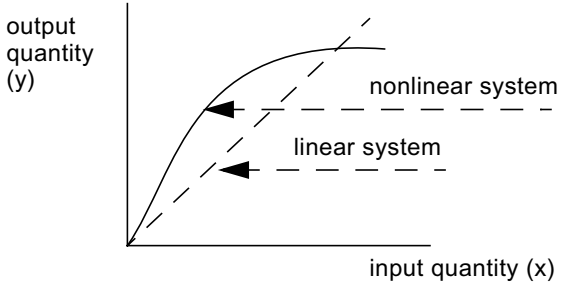


Figure 2.3 Linear and nonlinear system.

The nonlinearity of a system can be modeled in a number of ways. One way that allows easy calculation of spectral components is polynomial modeling, used throughout in this book. The output of the system modeled with a third-degree polynomial is written as

$$y = a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3, \tag{2.3}$$

where a_1 to a_3 are real valued nonlinearity coefficients at this stage of the analysis. The first term, a_1, describes the linear small-signal gain, whereas the a_2 and a_3 are the gain constants of quadratic (square-law) and cubic nonlinearities, introducing the curvature effects shown in Figure 2.3. In this chapter, the analysis is limited to third-degree, but up to fifth-degree effects will be discussed in Chapter 3.

The output of the nonlinear system can be calculated by substituting a single-tone sinewave (2.2), shown graphically in Figure 2.4(b), into (2.3). In the frequency domain, nonlinearity generates new spectral components shown in Figure 2.4(a) and Table 2.3. The output comprises not only the fundamental signal (ω_1), but also the second harmonic ($2\omega_1$) and dc (0) generated by a_2x^2 and the third harmonic ($3\omega_1$) generated by a_3x^3. This spectral regrowth, which will be discussed in more detail later, is not possible in linear systems. Figure 2.4(b) shows that, in nonlinear systems, the steady-state time domain output waveform is a distorted copy of the input waveform. Like spectral regrowth, this phenomenon is not possible in linear systems, in which the steady-state output signal is always identical in shape to the input (i.e., it can only be attenuated/amplified and/or phase-shifted).

Table 2.3

Amplitude of Spectral Components Generated by a Single-Tone Test and Nonlinearities Up to the Third Degree

dc	Fundamental	2nd Harmonic	3rd Harmonic
$(a_2/2)A^2$	$a_1A+(3a_3/4)\ A^3$	$(a_2/2)A^2$	$(a_3/4)A^3$

If the nonlinearity coefficients in (2.3) have real values, the system is considered nonlinear and memoryless, because the fundamental output signal is in phase with the input over the whole frequency range. If the

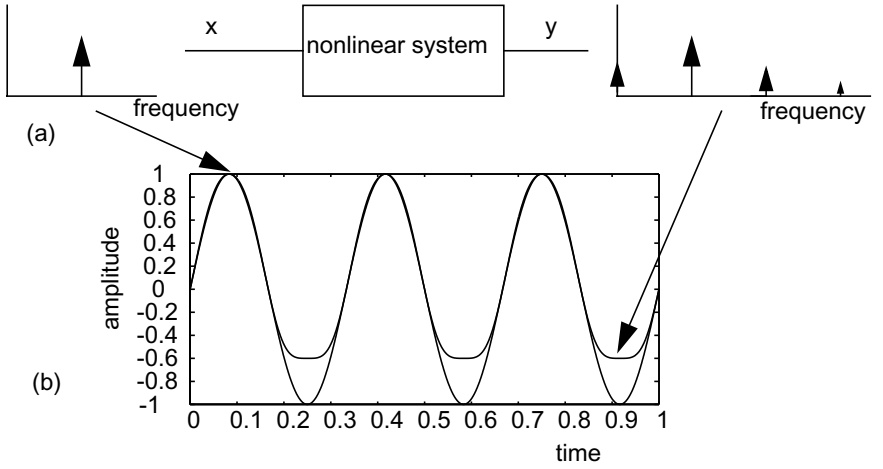


Figure 2.4 Nonlinear effects in frequency and time domains. (a) Input and output spectrums and (b) waveforms.

coefficients include a phase shift (which appears as a complex-valued coefficient), a constant, frequency-independent phase shift will exist between the input and output signals, thus modeling a nonlinear system with memory. Complex-valued coefficients are normally used in narrowband behavioral models, as will be shown later. Here it suffices to note that memory causes phase shift in nonlinear systems in much the same way as in linear systems.

2.1.3 Common Measures of Nonlinearity

We now look at the effects of nonlinearity as a function of signal amplitude. As noted earlier, new signal components occur at the dc, fundamental, second, and third harmonics. The fundamental signal consists of the linear term a_1A and the third-order term $(3a_3/4)$A^3, while the third harmonic only comprises the third-order term. The dc and second harmonic terms are equal in amplitude and are both caused by the second power term $(a_2/2)$A^2. Figure 2.5 presents the spectral components at the output as a function of input signal level, obtained from a polynomial system (2.3) for a single-tone sinusoidal input (2.2). As seen from Table 2.3, the second and third harmonics increase to the power of two and three of the input amplitude. The fundamental signal, however, increases to the power of one at low signal levels, but at higher values, the cubic nonlinearity (or any

odd-degree nonlinearity in general) starts to modify the linear behavior of the fundamental signal. This means that the nonlinearity of the system can be considered in two ways: either a generation of new spectral components and/or an amplitude-dependent gain of the fundamental signal gain.

This gives two common measures for nonlinearity: 1-dB compression point P_{1dB} where the large-signal gain has dropped 1 dB, and intercept points (P_{IIP3}), where the extrapolated linear and distortion products cross. By using a third-degree polynomial amplifier model (2.3) with negative a_3 and single-tone test (2.2) for calculating the compression point and two-tone test (2.7) for IIP3, we get the common approximation stating that $P_{1dB} = P_{IIP3} - 10$ dB and that the IM3 level at the compression point is as high as –20 dBc.

Another widely used measure of nonlinearity is AM-AM and AM-PM conversions [1, 2]. These figures model the amplitude and phase of the fundamental signal with increasing input amplitude. The linear and third-order spectral components of a fundamental signal are shown separately in Figure 2.6 at a certain amplitude value. Due to the third power dependency of the upper vectors, the fundamental signal is increasingly modified as the signal amplitude increases. Figure 2.6(a) presents the situation already depicted in Figure 2.5. The values of a_1 and a_3 are real and have opposite signs, producing amplitude compression at high amplitude values. The second plot, Figure 2.6(b), presents the opposite situation in which a_1 and a_3 are both real and either positive or negative, resulting in AM-AM gain

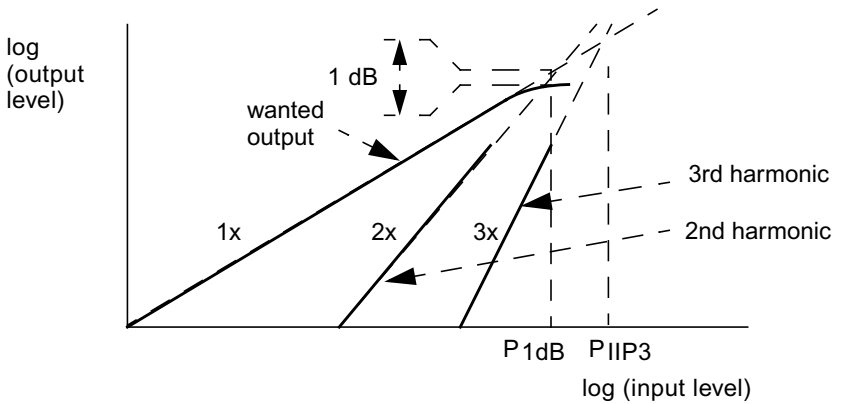


Figure 2.5 Illustration of nonlinear effects. The wanted (fundamental) output begins to change from its linear 1:1 slope at high amplitude levels and the generated spectral components increase as a function of signal amplitude.

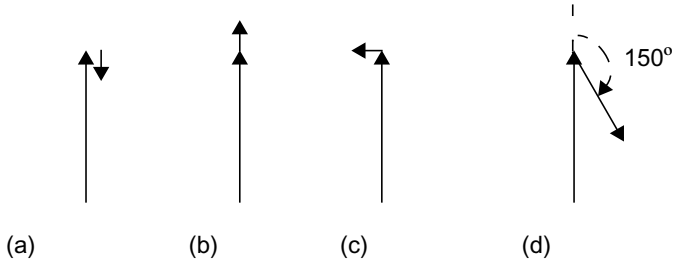


Figure 2.6 Amplitude and phase conversions caused by third-order distortion. (a) AM-AM compression, (b) AM-AM expansion, (c) AM-PM, and (d) the situation shown next in Fig. 2.7.

expansion. In the third plot, Figure 2.6(c), a_1 and a_3 display a phase difference that deviates from 0º or 180º, thereby producing an AM-PM conversion. Note that this combination of AM-AM and AM-PM cannot be predicted using a power series with real coefficients, but we need to have a complex value for a_3 in the phasor calculations.

This reasoning can be extended to higher order distortion analysis, as well. If, for example the third-order term is in-phase and fifth-order term is in an opposite phase with the linear term, we have a response where the gain first expands due to cubic nonlinearity and then compresses due to fifth-degree nonlinearity, when the signal level is increased.

We now consider the case shown in Figure 2.6(d), where the magnitude and phase of a_3 are 0.1 and 150º, respectively, while the corresponding values for a_1 are 1 and 0º. Figure 2.7 shows AM-PM as a function of fundamental gain compression (AM-AM), with a value of approximately 3.5º at the 1-dB compression point. It must be emphasized here that a system operating at 1-dB compression is already heavily nonlinear. Linearity requirements are so demanding nowadays that amplifiers are backed-off well below the 1-dB compression point, and their AM-PM may be as low as 1º or 2º at full power and approach zero with decreasing power.

The value of AM-PM is very small, so it is a difficult parameter to measure accurately. Phase changes in the fundamental signal introduced by AM-PM depend on signal amplitude, and very high values are needed to make a visible effect. The same observation holds for AM-AM. The problem with using amplitude conversions as a figure of merit for nonlinearity is that they measure nonlinearity on the basis of the fundamental signal, which comprises a strong linear term. Since nonlinear effects in the fundamental are small, the measurement of AM-AM and AM-PM is highly sensitive to measurement errors.

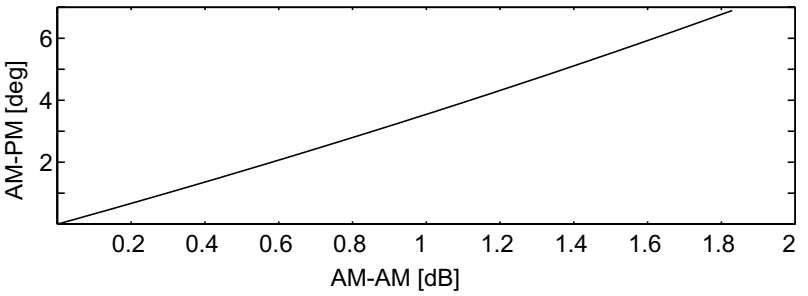


Figure 2.7 AM-PM of a polynomial system as a function of AM-AM. From [3].

Throughout this book, nonlinearity is considered by studying the behavior of generated new spectral components. Using the polynomial input-output model, the same information about nonlinearity (a_3) can be seen both from amplitude conversions and the third harmonic component (or third-order intermodulation term IM3 in the case of a two-tone test). Technically, it is easier and more robust to measure and analyze the behavior of distortion tones than AM-AM, in which the nonlinear effects appear only as small variations on top of a strong fundamental signal.

2.2 Calculating Spectrums in Nonlinear Systems

Integral transforms like Fourier or Laplace transform can be used to simplify the analysis of linear systems. With some care, their use can be extended to nonlinear or time-varying systems as well.

It is well known that the time-domain response $y(t)$ of a linear circuit is the convolution of the impulse response $h(t)$ and the input signal $x(t)$, as shown in (2.4). In the frequency domain this converts to a multiplication of the frequency response $H(j\omega)$ and the signal spectrum $X(j\omega)$.

$$y(t) = h(t)*x(t) \leftrightarrow \quad Y(j\omega) = H(j\omega) \cdot X(j\omega), \tag{2.4}$$

where the convolution (*) is calculated with (2.5). A graphical interpretation of convolution (used later in Figure 2.8) is that for each value of t, we reverse the time axis of $x(\tau)$, shift it by the amount of t, and then integrate the product of $h(\tau)$ and time-reversed and shifted $x(\tau)$ over all previous values of t, and store the result in place of $y(t)$.

$$h(t)*x(t) = \int_{-\infty}^{\infty} h(\tau)x(t-\tau)d\tau \tag{2.5}$$

For nonlinear systems the convolution operates the other way around: a time domain multiplication of two signals corresponds to frequency domain convolution of their spectrums.

$$y(t) = x(t) \cdot x(t) \leftrightarrow \quad Y(j\omega) = X(j\omega)*X(j\omega) \tag{2.6}$$

Similarly, the spectrum of $y(t)=x(t)^N$ is obtained simply by taking an N-fold convolution of $X(j\omega)$ with itself. It may sound overly academic to calculate the spectrum of a nonlinear system as a multiple convolution of the linear signal spectrum, but in fact (2.6) is an extremely handy and effective way of calculating the line spectrum of a multitone signal numerically (see [4]), and either a symbolic or graphical convolution illustrated in Figure 2.8 is a rigorous way of obtaining all the possible mixing results falling to a given distortion tone. When performed with complex numbers, the convolution also preserves the phase information of the tones.

As an example, the output spectrum of a two-tone test signal in quadratic nonlinearity can be calculated as follows. The two-tone signal is given by

$$\begin{aligned} x &= A_1 \cdot \cos(\omega_1 t + \phi_1) + A_2 \cdot \cos(\omega_2 t + \phi_2) \\ &= \frac{A_1 e^{j\phi_1}}{2} \cdot e^{j\omega_1 t} + \frac{A_1 e^{-j\phi_1}}{2} \cdot e^{-j\omega_1 t} \\ &\quad + \frac{A_2 e^{j\phi_2}}{2} \cdot e^{j\omega_2 t} + \frac{A_2 e^{-j\phi_2}}{2} \cdot e^{-j\omega_2 t} \end{aligned} \tag{2.7}$$

that is presented in Figure 2.8(a) using a two-sided spectrum. The right-hand side of the plot represents the positive frequency axis, and A_1 and A_2 are now complex numbers containing both the amplitudes ($A_j/2$) and phases of lower (ω_1) and higher (ω_2) tones, respectively. Due to odd phase response of real systems, the phasors $\overline{A}_1$ and $\overline{A}_2$ of the negative frequencies on the left are complex conjugates of A_1 and A_2. Figure 2.8(b) is identical to Figure 2.8(a), whereas Figure 2.8(c) presents the original input spectrum with a reversed frequency axis: Positive frequencies are now on the left and negative frequencies on the right. Next, the reversed spectrum is slid from right to left and compared at all offsets to the original input in Figure

2.8(a). Figure 2.8(d) presents the situation at a single frequency offset, that corresponds to a single frequency in the output spectrum. Now we simply multiply all the aligning frequency pairs [shown with dashed line between Figure 2.8(a, d)] and place the sum of these products ($\overline{A}_1A_2+A_2\overline{A}_1$) as the amplitude (actually a phasor) of the generated tone. The frequency offset between Figure 2.8(a), (d) corresponds to the envelope frequency $f_2–f_1$ (also called the beat, video, or modulation frequency), but the other tones are generated similarly. For example, a frequency offset $2\omega_1$ [i.e., the origin of the spectrum Figure 2.8(d) aligns with frequency $2\omega_1$ in the original spectrum Figure 2.8(a)] causes the A_1 phasors in Figure 2.8(a), (d) to align, resulting in a second harmonic with amplitude $A_1{}^2$ in spectrum (e). Finally, Figure 2.8(e) presents the complete spectrum generated by squaring the two-tone signal in Figure 2.8(a). The procedure demonstrated in Figure 2.8 is known as spectral convolution.

Note that it is necessary to use a two-sided spectrum to calculate the amplitudes of the distortion tones using the spectral convolution. Hence, all amplitudes except the dc term include the term 1/2.

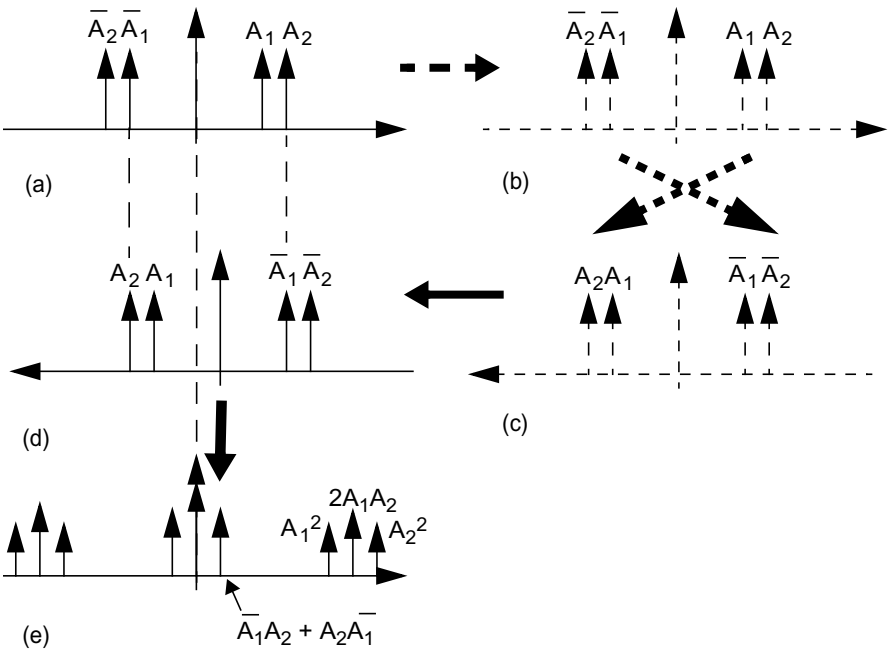


Figure 2.8 Spectral convolution. (a) The original and (b)-(c) flipped spectrum; (d) shows the flipped and shifted spectrum, and (e) is the final convolution result. Note that the phasors include the coefficient 1/2.

2.3 Memoryless Spectral Regrowth

This section discusses the spectral regrowth in a memoryless nonlinearity. A block presentation of a nonlinear system modeled by an input-output polynomial (2.3) is given in Figure 2.9, where the output is the sum of the first, second, and third powers y_1, y_2, and y_3 of the input signal, weighted by the nonlinearity coefficients a_1, a_2, and a_3, respectively. In phasor analysis, the coefficients may be complex to model the phase shift in the nonlinearities. The spectrums in the intermediate points A and B can be calculated as a two- and three-fold convolution of the two-sided input spectrum, respectively. As an example, the line spectrum of a squared two-tone signal in point A is shown in Figure 2.8(e).

This polynomial system is usually analyzed by assuming that $x(t)$ is a nondistorted two-tone signal. In this case, the linear term a_1x just amplifies the fundamental tones at ω_1 and ω_2 ($\omega_2>\omega_1$). The quadratic nonlinearity a_2x^2 rectifies the signal down to dc band to frequencies 0 Hz (dc) and $\omega_2-\omega_1$. It also generates the second harmonic band consisting of tones at $2\omega_1$, $2\omega_2$ and $\omega_1+\omega_2$, called the lower and higher second harmonic and the sum frequency, respectively. Similarly, the cubic nonlinearity a_3x^3 generates lower and higher IM3 at $2\omega_1-\omega_2$, and $2\omega_2-\omega_1$ and the compression/expansion terms (AM-AM) on top of the fundamental tones ω_1 and ω_2, all appearing in the fundamental signal band. It also generates the entire third harmonic band consisting of tones at $3\omega_1$, $2\omega_1+\omega_2$, $\omega_1+2\omega_2$, and $3\omega_2$, called the lower third harmonic, the lower and higher sum frequencies and the higher third harmonic, respectively. These tones are illustrated in the line spectrum shown in Figure 2.10.

Distortion tones are classified as harmonic (HD) and intermodulation (IM) distortion, where the harmonic distortion is simply an integer multiple of one of the input tones and IM tones appear at frequencies

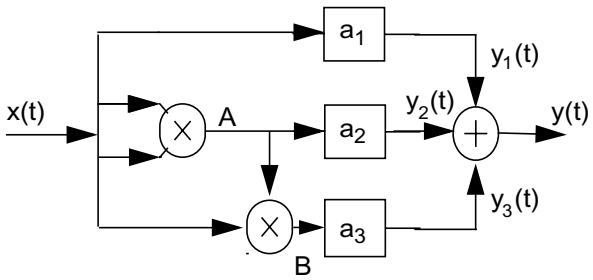


Figure 2.9 Block presentation of a memoryless system up to the third degree.

$$K\omega_1+L\omega_2, \tag{2.8}$$

where K and L are positive or negative integers. Another, more practical classification is based on the grouping of the tones: in RF applications, the dc, fundamental, second and third harmonic bands are far from each other and quite easily filtered separately, if needed. However, the IM3 distortion appearing in the fundamental band cannot usually be separated from the desired linear term.

The third and the most important classification is based on the order of the distortion product, which in short means the number of fundamental tones that need to be multiplied to make a distortion product of a given order. In a two-tone excitation in Figure 2.10, the fundamental tones ω_1 and ω_2 are first-order signals, while dc (0 Hz), envelope ω_2–ω_1, second harmonics $2\omega_1$ and $2\omega_2$, and the sum frequency ω_1+ω_2 are second-order signals. These build up the dc and second harmonic bands. Similarly, third-order signal components lay in the fundamental ($2\omega_1$–ω_2, ω_1, ω_2, $2\omega_2$–ω_1) and third harmonic bands ($3\omega_1$, $2\omega_1$+ω_2, ω_1+$2\omega_2$, $3\omega_2$). The amplitudes of the Nth-order tones always are proportional to A^N, where A is the amplitude of the fundamental tone(s).

Using the notations of (2.8), the order N is sometimes written as N= $|K|+|L|$. However, this rule breaks down when higher order tones fall on top of the lower order ones. As an example, look at the fifth-order compression term (2.9) below that appears at frequency $1\omega_1+0\omega_2$ but still is of the fifth order.

$$(A^5/32)\cdot e^{j\omega_1}\cdot e^{j\omega_1}\cdot e^{j\omega_2}\cdot e^{-j\omega_1}\cdot e^{-j\omega_2} \tag{2.9}$$

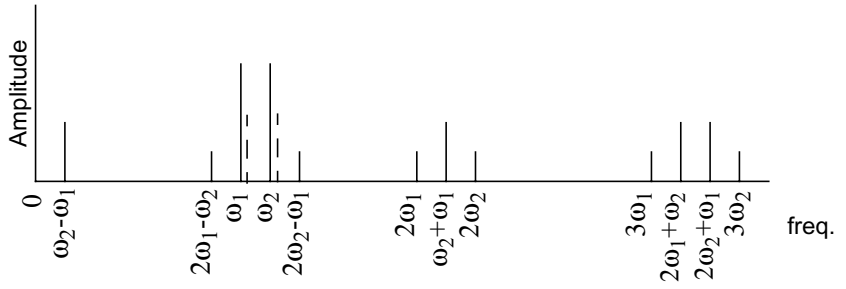


Figure 2.10 Spectral regrowth of a two-tone signal. AM-AM is shown as a dashed line next to fundamental tones.

Then what is the difference between the order of distortion and the degree of nonlinearity? So far the input signal has always consisted of first-order signals only, and the things have been simple: the first-degree term a_1x in (2.3) generates first-order tones, the second-degree (quadratic) term a_2x^2 second-order tones, and the third-degree (cubic) term third-order tones. However, the case is not so simple any more, if the input signal is already distorted, which is the typical case inside a real amplifier. A second-degree nonlinearity x^2 essentially makes a product x_1x_2, where the x_1 and x_2 are certain input tones. These need not be the same, and their order may already be higher than one. For example, multiplying the fundamental tone ω_1 with a second harmonic $2\omega_2$ inside a second-degree nonlinearity generates two third-order tones at $2\omega_2-\omega_1$ and $2\omega_2+\omega_1$. Hence, the order of the output tone is the sum of the orders of the input tones x_1 and x_2. In one extreme, a purely quadratic (second-degree) nonlinearity is capable of generating any order of distortion, if the distorted output is always fed back to the input.

To summarize, the term order is a property of the final distortion product, and it is related to the amplitude dependency and frequency of the distortion tone. The term degree is a property of the nonlinear device, defining the shape of the nonlinearity. The order of the distortion caused by an *N*th-degree nonlinearity depends both on the degree of the nonlinearity and the order of the input signals. In an *N*th-degree nonlinearity, *N* tones are multiplied, and the total order is the sum of the orders of these *N* tones.

This is illustrated in Table 2.4, where the amplitudes of all the tones generated by a third-degree polynomial are shown in a case where the input signal is a sum of the fundamental two-tone signal with phasors A_1 and A_2 and the second-order distortion tones DC, E, H_{11}, H_{12}, and H_{22} at frequencies 0, $\omega_2-\omega_1$, $2\omega_1$, $\omega_1+\omega_2$, and $2\omega_2$, respectively.We see that in this case, also the second-degree (quadratic) nonlinearity a_2x^2 can generate third-order distortion appearing at the fundamental and third harmonic bands.

Note that contrary to most presentations in textbooks, Table 2.4 also contains the phase information and allows the calculation in a case of unequal tone amplitudes as well. The table gives the amplitudes and phases for a one-sided spectrum (i.e., they are directly the amplitudes of the sinusoids), and to make a two-sided spectrum, simply divide all but dc by 2 and substitute the complex conjugates of the positive phasors to the negative frequencies. This table is already quite difficult to build analytically, but the encircled terms are easily found by drawing the spectrum of the second-order tones and convolving it graphically with a two-tone spectrum.

Table 2.4

Spectral components generated in a third-degree polynomial nonlinearity $y = a_1x + a_2x^2 + a_3x^3$ for a sum of two-tone signal phasors A_1 and A_2 and second-order distortion phasors E, H_{11}, H_{22}, and H_{12} at $\omega_2-\omega_1$, $2\omega_1$, $2\omega_2$, and $\omega_1 + \omega_2$, respectively. The terms inside the boxes present the third-order results generated from first and second-order signals in the input.

Frequency	Name	a_1 x	$(a_2 / 2)$ x	$(a_3 / 4)$ x
0	DC	DC	$\overline{A}_1A_1+\overline{A}_2A_2$	
$\omega_2-\omega_1$	envelope	E	$2\overline{A}_1A_2$	
$2\omega_1-\omega_2$	IM3L		$\overline{E}A_1+H_{11}\overline{A}_2$	$3A_1{}^2\overline{A}_2$
ω_1	FUNDL	A_1	$\overline{E}A_2+H_{11}\overline{A}_1+$ $2DCA_1+H_{12}\overline{A}_2$	$6A_1A_2\overline{A}_2+3A_1A_1\overline{A}_1$
ω_2	FUNDH	A_2	$EA_1+H_{22}\overline{A}_2+$ $2DCA_2+H_{12}\overline{A}_1$	$6A_1\overline{A}_1A_2+3A_2A_2\overline{A}_2$
$2\omega_2-\omega_1$	IM3L		$EA_2+H_{22}\overline{\mathrm{A}}_1$	$3\overline{A}_1A_2{}^2$
$2\omega_1$	2HL	H_{11}	$A_1{}^2$	
$\omega_1 + \omega_2$	2SUM	H_{12}	$2A_1A_2$	
$2\omega_2$	2HH	H_{22}	$A_2{}^2$	
$3\omega_1$	3HL		$H_{11}A_1$	$A_1{}^3$
$2\omega_1 + \omega_2$	3SUML		$H_{11}A_2+H_{12}A_1$	$3A_1{}^2A_2$
$2\omega_2 + \omega_1$	3SUMH		$H_{22}A_1+H_{12}A_2$	$3A_1A_2{}^2$
$3\omega_2$	3HH		$H_{22}A_2$	$A_2{}^3$

2.4 Signal Bandwidth Dependent Nonlinear Effects

Section 2.1 described the classification of electrical systems into linear and nonlinear systems with and without memory. This classification is presented graphically in Figure 2.11, in which the overlapping segment between two areas represents nonlinear systems with memory. This segment is further subdivided into two sections. The upper section represents a narrowband system, where the transfer function is dependent on the center frequency of the system only, while the lower section represents a system that is also affected by the bandwidth of the input signal. Since all practical systems are more or less affected by signal bandwidth, the upper section is referred to as a narrowband approximation of a real, bandwidth-dependent system. In this book, bandwidth-dependent effects are called memory effects.

The narrowband single-tone signal used in Section 2.1 is insufficient for the characterization of memory effects. Instead, these effects can be investigated by applying a two-tone input signal with variable tone spacing. The alternative would be to use a real, digitally modulated signal, but it would yield less insight in the operation of the analyzed system, as will be seen later on. In addition, using a digitally modulated signal for the calculation of generated spectral components necessitates a time domain analysis tool with a Fourier transformation. The use of a sinusoidal input signal circumvents this problem, because spectral components can be calculated analytically.

This book studies the effects of variable tone spacing in detail to characterize bandwidth-dependent effects. Applying a two-tone signal to a third-degree polynomial system (2.3) results in the following two

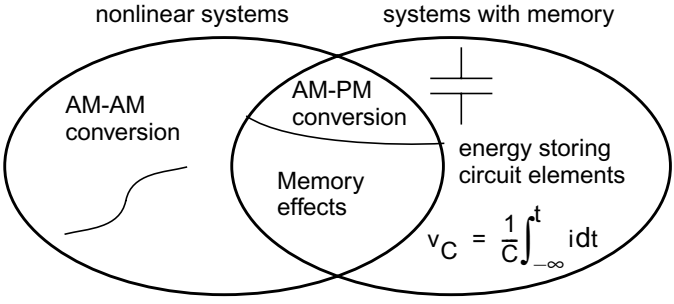


Figure 2.11 Definition of memory effects used in this book. From [3].

conclusions concerning IM3 signals at the output: first, they are not functions of tone spacing and, second, their amplitude increases exactly to the third power of the input amplitude. This is shown by the last column and third row in Table 2.4. The equation for the IM3L (lower IM3) component is proportional to the power of three while being independent of signal bandwidth. However, a comparison between the polynomially modeled and actual phases of the IM3L as a function of tone difference in a two-tone signal is sketched in Figure 2.12, where large differences can be observed between the two. The real phase (and amplitude) of the IM3 may deviate at low and high tone spacings (or modulation frequencies), indicating the existence of signal bandwidth-dependent nonlinear effects with memory, as marked by the lower overlapping area in Figure 2.11. This book refers to such effects as memory effects, and distinguishes between two distinct types: electrothermal memory effects, which typically appear at low modulation frequencies (below 100 kHz), and electrical memory effects appearing above MHz modulation frequencies.

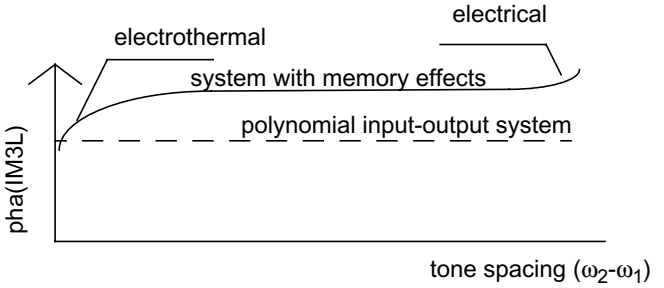


Figure 2.12 Phase of the IM3 component of a system with (solid line) and without (dashed line) memory effects. © IEEE 2001 [5].

The fundamental output of a two-tone input is also modified by a third-degree nonlinearity, shown in Figure 2.10 and Table 2.4. As a result, the two-tone signals are also affected by the amplitude and phase conversions. It then follows that memory effects can be characterized as changes in these conversions produced by a varying two-tone input [6]. Unfortunately, a two-tone input is hampered by the same drawbacks as a one-tone input. Strong linear signals at the fundamental make nonlinear effects difficult to measure. This is particularly important in the characterization of memory effects, which are usually very weak compared to linear signals. Therefore, the analysis of intermodulation components is the most practical starting point for the exploration of memory effects.

2.5 Analysis of Nonlinear Systems

Most nonlinear analysis/simulation methods operate either fully or partially in the time domain. Standard transient analysis based on numerical solving of nonlinear differential equations is an example of the former, and widely used harmonic balance method presents the latter. Here the passive components are modeled in the frequency domain, but still the responses of the nonlinear components are solved in time domain, and outputs and excitations are pumped back and forth between time and frequency domain using the discrete Fourier transform. Transient analysis can handle any form of input signal or even autonomous circuits (oscillators), but it suffers from ineffective modeling of distributed components and long-lasting initial transients that need to settle before the steady-state spectrum can be calculated. In the harmonic balance the signal is necessarily modeled by just a few sinusoids, but the initial transient is bypassed and more accurate frequency domain models can be used for passive components. An in-depth comparison of the basic simulation algorithms can be found in [7].

The Volterra analysis technique used in this book is calculated entirely in the frequency domain, building higher order responses recursively using lower order results. Hence, no iteration is needed and it is a very quick and RF-oriented analysis method. What is even more important in studying the memory effects is that it *can separate the sources of distortion* exactly in the same way engineers are accustomed to doing in noise simulations: The dominant contributions can be listed, and the designer can attack them first. That kind of information is very valuable for design optimization, but usually impossible to derive from transient or harmonic balance simulations that usually display only the total amount of distortion.

In the Volterra analysis, some simplifications and assumptions are made, though. The first simplification is that like in harmonic balance, only the sinusoidal steady-state response of a single or two-tone excitation is calculated. Second, the nonlinearities of the system are modeled polynomially (2.3). Using these assumptions, we may apply the Volterra method for calculating the output of a nonlinear system, which can give either numerical or analytical results for the distortion components.

The Volterra analysis is reviewed in Section 2.5.1, while Section 2.5.2 describes the direct or nonlinear current method for calculating nonlinear responses. Section 2.5.3, in turn, compares two Volterra modeling methods, the first of which provides merely input-output information, whereas the other one offers a true insight into the operation of the system. The latter method will be used throughout this book for its visualization and optimization benefits. More background information can be found in Appendix A.

2.5.1 Volterra Series Analysis

Volterra analysis can be considered a nonlinear extension of linear ac analysis, and its main difference compared to the often-used power series analysis is that it contains also the phase information of the transfer functions. It is often calculated symbolically [8-11], in which case the transfer functions describing the amplitude and phase of the distortion tones as functions of input signals are derived. These transfer functions are illustrated in Figure 2.13(a), where H_1 is the linear (small-signal) transfer function, H_2 is the second-order transfer function (producing all the second-order tones in a two-tone test), and so forth; the total output $y(t)$ is a sum of all these transfer functions applied to the input signal $x(t)$ [8,12,13].

The difference between linear and Volterra analysis is further illustrated in Figure 2.13(b). Linear small-signal ac analysis models the x-y input-output characteristic of a circuit element with its first derivative in the operating bias point. In Volterra analysis, the actual shape of the I-V or Q-V curve is modeled by a best fit, low-degree polynomial function of the controlling voltage, and the higher-degree coefficients of the polynomial are used to calculate the distortion components.

The output of the second-order Volterra kernel for the one-tone sinewave (2.2) is derived in Appendix A for interested readers. However, since the spectral components at the output can be calculated using the direct calculation method explained in the next section, an in-depth understanding of Volterra kernels is beyond the scope of this book. In a

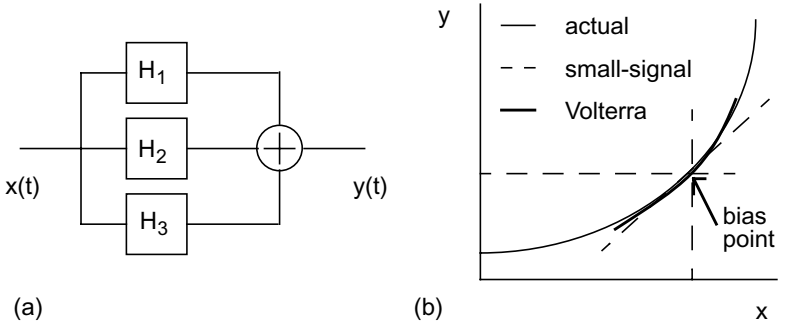


Figure 2.13 (a) Schematic representation of a system characterized by a Volterra series and (b) comparison between small-signal and Volterra series analysis. From [3].

fully numerical form, Volterra analysis has been implemented for example, in SPICE [14] and Voltaire XL [15] circuit simulators.

A word of warning concerning the Volterra analysis is needed. First, the polynomial models are notorious for the fact that their response explodes outside the fitting range - hence, the model is only locally fitted around the desired bias point and applicable over a certain amplitude range. The applicable range depends on the fitting range of the polynomial (as anything can happen outside the fitting range), the nonlinearity of the device, and the degree of the modeling polynomial. As the degree of the model always needs to be limited at some rather low degree, some truncation error between the actual and the modeled response exists. The effects of this truncation error are discussed in some more detail in Appendix B.

Figure 2.13 and the block diagram in Figure 2.9 model only an input-output nonlinearity. Due to intentional or nonintentional feedback, most of the controlling nodes in real amplifiers also contain distortion components and not just the linear contribution. This causes multiple mixing, as already pointed out in Table 2.4. For example, the second harmonic may mix with the linear term in a second-degree nonlinearity and generate IM3. These effects can be taken into account as well, by keeping track of the order of the calculated result. If v_1 contains all linear and v_2 contains all second-order voltage phasors and so forth, $v_1+v_2+v_3+...$ can now be substituted into the polynomial as shown in (2.10). After expanding this we can collect the terms of a given order on separate rows, and the last complete row in (2.10) shows that the third-order output current i actually consists of three terms: input third-order distortion v_3 (if present) multiplied by the linear gain a_1, input linear signal v_1 distorted in the cubic nonlinearity a_3x^3, and finally, a mixing result of linear and second-order input signals v_1 and v_2, generated in the quadratic nonlinearity a_2x^2. Also, higher order terms like v_2^2 or v_3v_2 are generated, but they are ignored in this analysis.

$$\begin{aligned} i &= a_1(v_1 + v_2 + ...) + a_2(v_1 + v_2 + ...)^2 + a_3 \cdot (v_1 + v_2 + ...)^3 \\ &= a_1 v_1 \\ &\quad + a_1 v_2 + a_2 v_1^2 \\ &\quad + a_1 v_3 + a_3 v_1^3 + 2a_2 v_1 v_2 \\ &\quad + ... \end{aligned} \tag{2.10}$$

2.5.2 Direct Calculation of Nonlinear Responses

A thorough study of the Volterra series can be found in [11, 13], while this book focuses on the frequency domain analysis only. Furthermore, rather than determining all third-order products, researchers usually concentrate on IM3 responses, which makes it impractical to derive the general form *N*th-order Volterra kernels. Instead, the direct method [11], also known as the nonlinear current method, can be employed to calculate only the desired signal components.

The direct method is based on modeling the nonlinear I-V and Q-V characteristics given by the polynomial functions in Figure 2.14 with a parallel combination of a linear element and a nonlinear current source, the current of which depends on the polynomial coefficients and the controlling voltages. This is illustrated in Figure 2.14, where K_{1G} to K_{3G} model the I-V and K_{1C} to K_{3C} the Q-V characteristic curves of a nonlinear conductance and capacitance, respectively. The linear terms K_{1G} and K_{1C} are modeled by a linear conductance and capacitance, and the higher degree nonlinearities are modeled by nonlinear voltage-controlled current sources. Furthermore, as the Q-V polynomial models the charge, it has to be differentiated with time to get the ac current. Note that the control voltage v may include also distortion voltages, which results in multiple mixing mechanisms, as illustrated in (2.10) and Table 2.4.

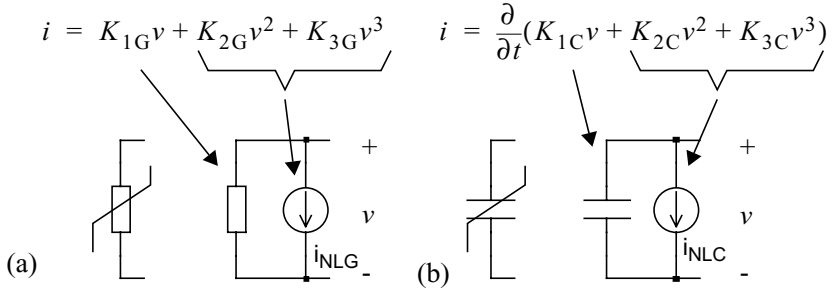


Figure 2.14 Equivalent models for nonlinear (a) conductance and (b) capacitance.

The procedure for the calculation of the response to a two-tone signal can be summarized as follows:

1. Evaluate the fundamental (first-order) node voltages using linear ac analysis for both tones.
2. For each nonlinear component, evaluate the second-order distortion currents using the fundamental voltage amplitudes. These will appear at five sum and difference frequencies: dc, envelope $\omega_2-\omega_1$, second harmonics $2\omega_2$, $2\omega_1$, and sum frequency $\omega_2+\omega_1$.
3. Use these distortion currents to calculate the second-order distortion voltages in each node using the ac analysis. Note that the distortion voltages are deterministic signals and they are summed as vectors, not as powers as in noise analysis.
4. Using the first- and second-order voltages, calculate the third-order distortion currents in the nonlinear components. These will appear at eight frequencies, two of which are IM3 signals.
5. Perform the ac analysis again at the frequencies of the third-order distortion currents to find the third-order node voltages.

In short, the linear node voltages are solved first, using small-signal analysis. Then, nonlinear analysis is started by modeling the nonlinearities by current sources and short (open)-circuiting the linear voltage (current) sources. Using linear analysis again, the second-order voltage responses of the distortion currents are calculated, and the procedure can be repeated all the way to higher order responses. An example of the direct calculation method will be given later on in this chapter.

Since the nonlinearities of the circuit elements are modeled by current sources, they will be explained in more detail. Each nonlinearity in the circuit is represented by a current source, which is placed in parallel with a linearized small signal element. The second-order current sources are calculated on the basis of the two-tone test signal, and the values of the second-order current source (one-sided) amplitudes at the envelope $\omega_1-\omega_2$ the second harmonic frequencies $2\omega_1$ are given in Table 2.5 and those for the IM3 results in Table 2.6. Note that Table 2.6 from [11] does not give the AM-AM term of the fundamental tones or third harmonics, but these can be derived using Table 2.4.

In the tables, K_{2G} and K_{2C} are the second-degree conductive and capacitive nonlinearity coefficients and $V_{i,m,n}$ is the voltage of the node i at the frequency of $m\omega_1+n\omega_2$. For example, the third index is zero for responses at $2\omega_1$, because $2\omega_1 = 2\omega_1+0\omega_2$. Similarly, intermodulation responses always include both frequencies, as indicated by the second column, and a negative sign corresponds to a negative frequency, which is

necessary to make the envelope frequency $\omega_1-\omega_2$. If a negative frequency is needed, the voltage phasor for it is the complex conjugate of the phasor at the positive frequency.

The conductances are memoryless, but as $i=\mathrm{d}q/\mathrm{d}t$, the charge polynomial needs to be differentiated with respect to time, and this causes the $j\omega$ dependency in the nonlinear current of capacitances. Hence, capacitors cause very small distortion currents at low frequencies, but high currents at the harmonic bands. A two-dimensional conductance is controlled by voltages v_i and v_j (e.g., v_be and v_ce). Here, both controlling voltages need to have one-dimensional polynomials of their own, but there are also terms consisting of the cross-products of voltages on both ports. These additional cross-terms are listed in the tables.

Note again that the phase of negative frequency components is opposite to positive frequencies (i.e., $V_{\mathrm{i},-1,0} = \overline{V}_{\mathrm{i},1,0}$). Normal rules of complex arithmetic apply, and for example, c^2 is still a complex number with twice the phase angle and frequency of c, while $|c|^2 = c\overline{c}$ is a scalar real number at dc. Some care is needed in calculating the responses of IM products consisting of both positive and negative frequencies, as terms v_in^3 and $v_\mathrm{in}^2\overline{v}_\mathrm{in} = |v_\mathrm{in}|^2 v_\mathrm{in}$ have different phase angles and frequencies, even though their amplitudes are exactly equal.

As seen from Table 2.6, the third-order signal components are not just functions of cubic nonlinearities, but they are also affected by the second-order voltages and quadratic nonlinearities. For example, the distortion current i_NL generated by a nonlinear conductance at the higher IM3 frequency $2\omega_2-\omega_1$ has the amplitude and phase given by

$$\begin{aligned} I_{\mathrm{NLG}(i,-1,2)} &= 3/4 \cdot K_{3\mathrm{G}} \cdot \overline{V_{i,1,0}} \cdot V_{i,0,1}^2 \\ &\quad + K_{2\mathrm{G}} \cdot V_{i,0,1} \cdot V_{i,-1,1} \\ &\quad + K_{2\mathrm{G}} \cdot \overline{V_{i,1,0}} \cdot V_{i,0,2} \end{aligned} \tag{2.11}$$

where the first row shows the effect of cubic nonlinearity and the last two rows show the up- and downconversion of the envelope ($V_{\mathrm{i},-1,1}$ at $\omega_2-\omega_1$) and second harmonic ($V_{\mathrm{i},0,2}$ at $2\omega_2$) tones, respectively, and $\overline{V}_{\mathrm{i},1,0}$ is the same as $V_{\mathrm{i},-1,0}$ given in Table 2.6. This is the same result found in Table 2.4 and illustrates again that if the controlling voltage of the nonlinear component is distorted, also a quadratic nonlinearity ($K_{2\mathrm{G}}v^2$ in this example) can generate third-order distortion – or in a more general way, lower degree nonlinearity can also generate higher order distortion.

Table 2.5

Second-Order Currents Caused by Second-Degree Nonlinearities in a Two-Tone Test [11]

Type of Nonlinearity	Nonlinear Current at Frequency $\omega_1+\omega_2$ or $\omega_2-\omega_1$	Nonlinear Current at Frequency $2\omega_1$
(Trans) Conductance	$K_{2G1} \cdot V_{i,1,0} \cdot V_{i,0,\pm1}$	$\frac{1}{2} \cdot K_{2G1} \cdot (V_{i,1,0})^2$
Capacitor	$j(\omega_1 \pm \omega_2) \cdot K_{2C} \cdot V_{i,1,0} \cdot V_{i,0,\pm1}$	$\frac{j2\omega_1}{2} \cdot K_{2C} \cdot (V_{i,1,0})^2$
Two-dimensional conductance (cross-terms only)	$1/2 \cdot K_{2G1G2} \cdot V_{i,1,0} \cdot V_{j,0,\pm1} + 1/2 \cdot K_{2G1G2} \cdot V_{i,0,\pm1} \cdot V_{j,1,0}$	$1/2 \cdot K_{2G1G2} \cdot V_{i,1,0} \cdot V_{j,1,0}$

Table 2.6

Third-Order Currents Caused by Second- and Third-Degree Nonlinearities in a Two-Tone Test [11]

Type of Nonlinearity	Nonlinear Current at Frequency $2\omega_1 \pm \omega_2$
(Trans) Conductance	$K_{2G1} \cdot V_{i,1,0} \cdot V_{i,1,\pm1} + K_{2G1} \cdot V_{i,0,\pm1} \cdot V_{i,2,0} + 3/4 \cdot K_{3G1} \cdot V_{i,1,0}^2 \cdot V_{i,0,\pm1}$
Capacitor	$(2j\omega_1 \pm j\omega_2) \cdot [K_{2C} \cdot V_{i,1,0} \cdot V_{i,1,\pm1} + K_{2C} \cdot V_{i,0,\pm1} \cdot V_{i,2,0} + 3/4 \cdot K_{3C} \cdot V_{i,1,0}^2 \cdot V_{i,0,\pm1}]$
Two-dimensional conductance (cross-terms only)	$1/2 \cdot K_{2G1G2} \cdot [V_{i,0,\pm1} \cdot V_{j,2,0} + V_{i,1,0} \cdot V_{j,1,\pm1} + V_{i,1,\pm1} \cdot V_{j,1,0} + V_{i,2,0} \cdot V_{j,0,\pm1}] + 1/4 \cdot K_{3G12G2} \cdot [2 \cdot V_{i,0,\pm1} \cdot V_{i,1,0} \cdot V_{j,1,0} + V_{i,1,0}^2 \cdot V_{j,0,\pm1}] + 1/4 \cdot K_{3G1G22} \cdot [2 \cdot V_{i,1,0} \cdot V_{j,0,\pm1} \cdot V_{j,1,0} + V_{i,0,\pm1} \cdot V_{j,1,0}^2]$

2.5.3 Two Volterra Modeling Approaches

A Volterra representation can be presented either as an input-output or a circuit-level description. Having already discussed the input-output model described by polynomial nonlinearity coefficients, we now look at the component-level model, which takes into account the nonlinearities of each circuit element. A comparison between the two models will be illustrated using the direct calculation method. Also, the existence of complex coefficients in the input-output Volterra model and amplitude conversions will be discussed.

Figure 2.15(a) presents a simple nonlinear circuit with memory. The first-order lowpass filter includes a linear series resistor and a nonlinear shunt capacitor, introducing nonlinear effects with memory.

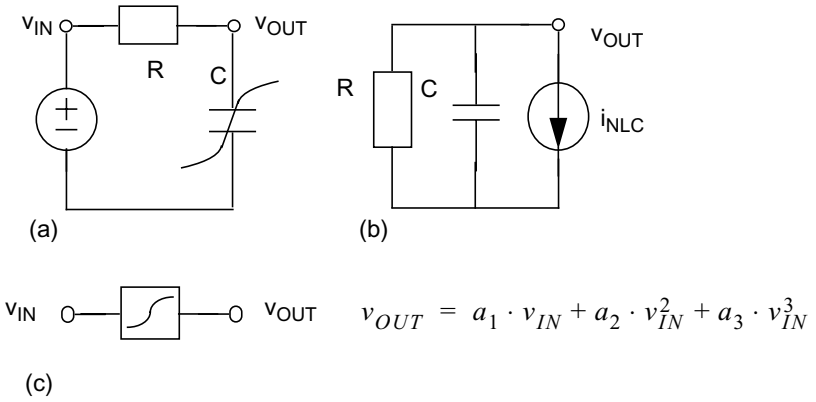


Figure 2.15 (a) Linearized first-order lowpass filter, and (b) circuit used for calculating nonlinear responses by a component-level Volterra approach. (c) The input-output Volterra modeling approach. From [3].

First, component-level calculations are applied using the direct method. The first-order, linearized transfer function is found to be

$$H_1(j\omega) = \frac{v_{\text{OUT}}}{v_{\text{IN}}} = \frac{1}{1 + j\omega CR}. \tag{2.12}$$

The circuit describing the nonlinear responses is shown in Figure 2.15(b). The linear voltage source is short-circuited and a nonlinear current

source is connected in parallel to the nonlinear element, in this case, the capacitance C. From Table 2.5, the value of the current source for the dc and second harmonic frequencies can be written as

$$\begin{aligned} i_{\mathrm{NLC}}(dc) &= 1/2 \cdot j(\omega_1 - \omega_1) \cdot K_{2\mathrm{C}} \cdot v_{\mathrm{OUT}}(\omega_1) \cdot v_{\mathrm{OUT}}(-\omega_1) \\ &= 0 \end{aligned} \tag{2.13}$$

and

$$i_{\mathrm{NLC}}(2\omega_1) = 1/2 \cdot j2\omega_1 \cdot K_{2\mathrm{C}} \cdot v_{\mathrm{OUT}}^2(\omega_1) \ , \tag{2.14}$$

where $K_{2\mathrm{C}}$ is the real-valued quadratic nonlinearity coefficient of the nonlinear charge. Note that the dc current caused by the nonlinear capacitor is zero, and no rectified dc voltage occur at the output. A second harmonic voltage exists, however, and is written as

$$v_{\mathrm{OUT}}(2\omega_1) = i_{\mathrm{NLC}}(2\omega_1) \cdot TF(2\omega_1), \tag{2.15}$$

where $TF(2\omega_1)$ is the value of transimpedance transfer function from the current source to the output node at frequency $2\omega_1$ – in this case simply the parallel impedance of linear R and C.

By combining (2.12), (2.14), and (2.15), the value of the second harmonic voltage in the output can be written as

$$\begin{aligned} v_{\mathrm{OUT}}(2\omega_1) &= i_{\mathrm{NLC}}(2\omega_1) \cdot TF(2\omega_1) \\ &= j \cdot \omega_1 \cdot K_{2\mathrm{C}} \cdot \left(\frac{v_{\mathrm{IN}}}{1 + j\omega_1 CR} \right)^2 \cdot \frac{R}{1 + j2\omega_1 RC} \end{aligned} \ . \tag{2.16}$$

It may be not be possible to minimize distortion by affecting the nonlinearity of the device (i_{NLC}), but usually the designer has some control over the harmonic terminal impedances (as well as in the dc bias impedance) and hence $TF(2\omega_1)$. This will be exploited later.

This simple component-level example contains all the information available from the system. Let us now try to compress it to a plain input-output model with fixed complex coefficients, illustrated in Figure 2.15(c) and commonly used in system simulations. Here, the output is written directly as a complex function of the input as follows

$$v_{OUT} = a_1 \cdot v_{IN} + a_2 \cdot v_{IN}^2 + a_3 \cdot v_{IN}^3 . \tag{2.17}$$

For a single-tone sine wave, the dc and second harmonic components at the output can be taken from Table 2.2 and are both equal to $(a_2/2)A^2$. A calculation using the component-level Volterra model shows that the dc term is zero, while the second harmonic is given by (2.16). This means that the input-output Volterra model with fixed coefficients fails to simultaneously model both the dc and the second harmonic component. Thus, if the scope is restricted to the dc component, we simply define a_2=0. However, if we are interested in the second harmonic, (2.16) and Table 2.2 allow us to write

$$a_2 = 2 \cdot j\omega_1 \cdot K_{2C} \cdot \left(\frac{R}{1 + j2\omega_1 RC}\right) \cdot \left(\frac{1}{1 + j\omega_1 CR}\right)^2 . \tag{2.18}$$

The same result is obtained for the second harmonic using both Volterra methods. This permits us to draw two important conclusions regarding the differences between the component-level and input-output methods. First, the value of the polynomial coefficient a_2 depends on frequency. If the frequency changes, its value has to be recalculated. The input-output Volterra modeling is accurate only at one frequency, which is why it is known as a narrowband approximation of a real system. Second, the value of a_2 can be complex to model the phase shift, even though the Q-V curvature is modeled by a real coefficient K_{2C}. Third, Volterra analysis includes phase information (ignored in the memoryless Taylor series), which makes it suitable for simulating high frequency effects such as AM-PM conversions as studied next using a cubic nonlinearity.

Tables 2.5 and 2.6 do not show the response at the fundamental, but the response for a plain third-degree nonlinearity can be obtained by raising a single-tone signal (2.2) to the third power and picking up the fundamental terms, which gives a one-sided amplitude of $(3/4)A^2\overline{A}$ (the same we get from Table 2.4 with A_2=0; or from Table 2.6 by setting $\omega_2=\omega_1$ and halving the amplitudes of the input tones). Here, the controlling voltage across the nonlinear C is v_{OUT} and hence

$$i_{NL3C}(\omega_1) = 3/4 \cdot j\omega_1 \cdot K_{3C} \cdot v_{OUT}^2(\omega_1) \cdot v_{OUT}(-\omega_1) . \tag{2.19}$$

The third-order compression/expansion of the fundamental output voltage can then be calculated as

$$v_{OUT3}(\omega_1) = i_{NL3C}(\omega_1) \cdot TF(\omega_1)$$
$$= 3/4 \cdot j\omega_1 \cdot K_{3C} \cdot \left(\frac{v_{IN}}{1 + j\omega_1 CR}\right)^2 \cdot \frac{\overline{v_{IN}}}{1 - j\omega_1 CR} \cdot \frac{R}{1 + j\omega_1 RC} . \quad (2.20)$$

Here, v_{OUT} is calculated using (2.12), and note that in the expression for the negative frequency $v_{OUT}(-\omega 1)$, the complex conjugate of phasor v_{IN} is used. The last term is again the transfer function from distortion current to output voltage, in this case again the parallel impedance of R and C, now calculated at ω_1. Comparing (2.20) with the input-output modeled compression/expansion taken from Table 2.2 (equal to $(3a_3/4)A^3$), the value of a_3 can be expressed as

$$a_3 = j\omega_1 \cdot K_{3C} \cdot \left(\frac{R}{1 + j\omega_1 RC}\right) \cdot \left(\frac{1}{1 + j\omega_1 CR}\right)^2 \cdot \left(\frac{1}{1 - j\omega_1 CR}\right) . \quad (2.21)$$

Like a_2, a_3 of the input-output model is a function of frequency and can be a complex value. The nonlinear blocks are often modeled by AM-AM and AM-PM conversions, which describe the gain and phase of the fundamental signal as a function of input amplitude. By taking the fundamental tone (including both the first- and third-order terms) from Table 2.2 and dividing the result by the input voltage phasor, the following equation can be written for amplitude conversions

$$\frac{v_{OUT}}{v_{IN}} = a_1 + 3/4 \cdot a_3 \cdot |v_{IN}|^2 . \quad (2.22)$$

The absolute value of (2.22) represents the AM-AM, and its phase represents the AM-PM. This is a mathematical formulation of the situation presented graphically in Figure 2.6. As illustrated by the figure, the phase difference between a_1 and a_3 determines the nature of the amplitude conversions. Finally, using (2.12) and (2.21) allows the conversions in this case study to be written as

$$\frac{v_{OUT}}{v_{IN}} = \frac{1}{1 + j\omega_1 CR} + 3/4 \cdot j\omega_1 \cdot K_{3C} \cdot \left(\frac{R}{1 + j\omega_1 RC}\right) \quad (2.23)$$
$$\cdot \left(\frac{1}{1 + j\omega_1 CR}\right)^2 \cdot \left(\frac{1}{1 - j\omega_1 CR}\right) \cdot |v_{IN}|^2$$

Equation (2.23) shows that due to the 90º phase shift of the third-order term, AM-PM conversion necessarily appears at high amplitudes. Equation (2.23) also shows that, similar to input-output Volterra modeling, models based on AM-AM and AM-PM conversions depend on the center frequency of the system. Since the conversions provide a narrowband approximation of a real bandwidth-dependent system, they fail to take into account the memory effects appearing inside the signal band.

This section has already given an example of the application of the direct method to distortion computations, illustrated the existence of complex coefficients in the input-output Volterra model, and provided some background information on AM-AM and AM-PM conversions. Finally, this section will also demonstrate the third-order distortion caused by quadratic distortion mechanisms that were neglected in (2.19) to simplify the analysis.

Third-order distortion is produced not only by K_{3C} but also by both the dc and second harmonic second-order voltages with the following mechanism. The controlling voltage of the quadratic nonlinearity contains both linear and second-order voltages v_1 and v_2, respectively. The quadratic nonlinearity K_{2C} acts now like a square law mixer, creating a third-order product $2v_1v_2$ as one term in the expansion of $(v_1+v_2)^2$. This is essentially a difference tone generated by a quadratic nonlinearity, and we can calculate its amplitude using the second row of Table 2.5 by replacing ω_1 with $2\omega_1$, ω_2 with $-\omega_1$, and using the minus sign in the equation:

$$i_{\mathrm{NLC}}(\omega_1) = j\omega_1 \cdot K_{2\mathrm{C}} \cdot v_{\mathrm{OUT}}(2\omega_1) \cdot v_{\mathrm{OUT}}(-\omega_1) \ . \tag{2.24}$$

Since no nonlinear dc component exists in the nonlinear capacitance, (2.24) includes only the second harmonic contribution. This shows that the third-order distortion is generated also by the cascaded second-order distortion mechanisms. Using (2.12) and (2.16), the compression term produced by this mechanism is given by

$$v_{\mathrm{OUT,\,K2}}(\omega_1) = -\omega_1^2 \cdot K_{2\mathrm{C}}^2 \cdot \left(\frac{R}{1 + j\omega_1 RC}\right) \cdot \left(\frac{R}{1 + j2\omega_1 RC}\right) \cdot \left(\frac{1}{1 - j\omega_1 CR}\right) \cdot \left(\frac{1}{1 + j\omega_1 CR}\right)^2 \cdot |v_{\mathrm{IN}}|^2 \cdot v_{\mathrm{IN}} \tag{2.25}$$

2.6 Summary

This chapter has introduced some important theoretical aspects of electrical circuits. Before a system can be analyzed, it has to be classified. The classical circuit theory divides circuits into linear or nonlinear systems that either exhibit memory or not. If a system is linear, the output is directly proportional to the input, no new spectral components can be generated, and the steady-state output waveform is identical in shape to the input waveform. Nonlinearity, however, means that the output is a nonlinear function of the input, so the gain of the system depends on the amplitude of the applied signal. Nonlinearity also introduces spectral regrowth and modifies the steady-state signal waveform.

In a memoryless system, the output is an instantaneous function of the input. Any change in the input signal occurs instantaneously at the output, therefore no phase difference exists between the input and output signals. Memory, however, makes the output also a function of previous input values. Thus, memory causes delays in transient signals, before the output settles to its steady-state value. This is caused by energy storing circuit elements. However, it is important to emphasize that memory itself does not modify the steady-state signal waveform, rather only introduces a phase shift between the input and the output.

We can look at the nonlinearity of a system in two ways. It can be seen as a modification of system gain (and phase) as a function of the applied signal amplitude, or as generation of new spectral components. The first view describes the fundamental signal that is modified by nonlinear effects. The higher the signal amplitude, the more the fundamental signal (its amplitude and phase) is modified by nonlinearities. The drawback of this classical way of looking at nonlinear effects is that telecommunication systems are becoming increasingly linear, making the characterization of nonlinearity difficult by using just the fundamental tone. The alternative involves the analysis of generated distortion components. These are easier to measure, because strong fundamental signals do not cause disturbances and also because the distortion components provide more information about the analyzed system.

RF power amplifiers are nonlinear circuits with memory. Nonlinear systems are often assumed to be unaffected by the input signal bandwidth, but this is not necessarily true for RF power amplifiers, as will be demonstrated in later parts of this book. Bandwidth-dependent effects, referred to as memory effects in this presentation, will be studied by looking at the amplitude and phase of IM3 components as a function of the tone spacing of the two-tone input signal.

Bandwidth-dependent nonlinear circuits with memory can be analyzed using the Volterra analysis. Here, the nonlinear I-V and Q-V characteristics are modeled as polynomial functions of the controlling voltages, where the first-degree terms are equivalent to the linear small-signal elements and the higher-degree terms are modeled by excess current sources parallel to the linear elements. The nonlinear response is calculated using the direct method, where the linear circuit is solved first to determine the fundamental node voltages. These voltages are inputs to the nonlinear current sources that model the nonlinearities of the circuit elements. Further, the currents transfer to node voltages, which are inputs to higher order responses, and so on. The key point here is the transformation from nonlinear current to nonlinear voltage, which is determined by the node (trans-) impedance at the frequency of the distortion current. Distortion can then be minimized not only by the fundamental impedance levels, but also by optimizing the out-of-band terminal impedances.

A Volterra representation can be regarded either as an input-output or a component-level description. The first corresponds to a polynomial between the input and output quantities, and the second comprises the actual schematic of the circuit elements whose nonlinearities are characterized by real valued nonlinearity coefficients. These two can be compared by making the component-level model first and then extracting the input-output model from it. The input-output model is a narrowband approximation of a real bandwidth-dependent system, which is accurate only at one frequency at a time. If the frequency of interest is changed, the coefficients of the polynomial have to be recalculated. This makes the input-output model an insufficient tool for the simulation of memory effects. Similar to the polynomial input-output model, AM-AM and AM-PM curves, widely used as a figure of merit of nonlinearity, are narrowband approximations of real bandwidth-dependent systems. Nevertheless, being well suited to the study of memory effects, the component-level Volterra method is the method of choice in this book.

One important property of the Volterra method is that the spectral components at the output can be expressed analytically, provided that the input signal is a sinusoid (or a sum of sinusoids). This is a unique property in polynomial modeling, which serves to reduce computational complexity, while providing an insight into the nonlinearity mechanisms of the system. This is one of the main reasons for the application of the polynomial Volterra approach throughout this book.

2.7 Key Points to Remember

1. Memory is caused by the storage of energy that has to be charged or discharged.
2. The nonlinearity of a system is easier to measure on the basis of generated spectra than on variations of the fundamental signal.
3. Bandwidth-dependent nonlinear effects are known as memory effects and can be analyzed using the component-level Volterra method.
4. The input-output Volterra method or AM-AM and AM-PM curves do not take memory effects into account.
5. The direct method can be used in distortion computations for nonlinear systems characterized by the Volterra series.
6. Analytical equations for distortion products can be calculated using the direct method that models the nonlinearities of the circuit elements using nonlinear current sources connected in parallel with a linearized, small-signal circuit elements.
7. Since the amount of device nonlinearities cannot be affected much, distortion is most effectively minimized by optimizing the impedances seen by the distortion current sources.

References

[1] Kenington, P. B., *High Linearity RF Amplifier Design,* Norwood, MA: Artech House, 2000.

[2] Jardon, A., and L. Vazquez, "A novel representation of AM-PM conversion," *Proc. 1995 IEEE International Symposium on Electromagnetic Compatibility,* pp. 401-405.

[3] Vuolevi, J., "Analysis, measurement and cancellation of the bandwidth and amplitude dependence of intermodulation distortion in RF power amplifiers," Doctoral thesis, University of Oulu, Oulu, Finland, 2001.

[4] Heiskanen, A., and T. Rahkonen, "5th order multi-tone Volterra simulator with component-level output," *Proc. 2002 IEEE International Symposium on Circuits and Systems*, Phoenix, AZ, 2002, pp. 591-594.

[5] Vuolevi, J., T. Rahkonen, and J. Manninen, "Measurement technique for characterizing memory effects in RF power amplifiers," *IEEE Trans. on Microwave Theory and Measurements*, Vol. 49, No. 8, 2001, pp. 1383-1389.

[6] Bösch, W., and G. Gatti, "Measurement and simulation of memory effects in predistortion linearizers," *IEEE Trans. on Microwave Theory and Techniques,* Vol. 37, No. 12, 1989, pp. 1885-1890.

[7] Miliozzi, P., et al., "Design of mixed-signal systems-on-a-chip," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 19, No. 12, 2000, pp. 1561-1571.

[8] Barrett, J., "The use of functionals in the analysis of nonlinear physical systems," *Journal of Electronics and Control*, Vol. 15, No. 6, 1957, pp. 567-615.

[9] Narayanan, S., "Application of Volterra series to intermodulation distortion analysis of transistor feedback amplifiers," *IEEE Trans. on Circuit Theory*, Vol. 17, No. 4, pp. 518-527.

[10] Maas, S., *Nonlinear Microwave Circuits,* Norwood, MA: Artech House, 1998.

[11] Wambacq, P., and W. Sansen, *Distortion Analysis of Analog Integrated Circuits,* Norwell, MA: Kluwer Academics, 1998.

[12] George, D., *Continuous Nonlinear Systems*, Technical Report No. 355, Research Laboratory of Electronics, M.I.T., 1959.

[13] Schetzen, M., *The Volterra and Wiener Theories of Nonlinear Systems*, New York: John Wiley & Sons, 1980.

[14] *HSPICE User's Manual Release 96.1*, Meta-Software Inc., 1996.

[15] *Microwave Office^{TM} User's Manual II*, Applied Wave Research, Inc., 2000.

Chapter 3

Memory Effects in RF Power Amplifiers

Memory effects, defined as bandwidth-dependent nonlinear effects, are the main topic in this chapter. RF power amplifiers play an important role in modern telecommunications, where opposite design goals make the performance optimization of amplifiers very difficult. Linearity was discussed in Chapter 2, while Section 3.1 investigates the quality of dc-to-RF conversion, which involves the primary problem with RF power amplifiers; namely, the trade-off between linearity and efficiency. To improve the trade-off, the amplifier can be designed to attain good efficiency at the expense of linearity. The linearity specification is then met by means of some external linearization technique. Unfortunately, memory effects cause a variation in intermodulation tones within the signal band. Although this may not dramatically decrease the linearity of the amplifier itself, it can considerably reduce the performance of the used linearization technique, thus deteriorating the trade-off between efficiency and linearity.

Section 3.1 defines amplifier efficiency, and Section 3.2 reviews the most common linearization techniques and discusses the consequences of memory effects in them. Next, Sections 3.3, 3.4, and 3.5 concentrate on different types of memory effects in RF power amplifiers, such as making a difference between electrical memory effects caused by nonconstant impedances and thermal memory effects caused by dynamic self-heating. Section 3.5 introduces the topic of amplitude-dependent memory effects that arise at moderate signal amplitudes.

3.1 Efficiency

Efficiency in power amplifiers describes the part of dc power that is converted to RF power and can be expressed as follows:

$$\eta = \frac{P_{OUT}}{P_{dc}}, \tag{3.1}$$

where P_{OUT} is the output RF power and P_{dc} is the power taken from the dc source. Power-added efficiency (PAE), however, takes the power of the input signal into account and can be expressed by

$$\text{PAE} = \frac{P_{OUT} - P_{IN}}{P_{dc}} = \eta \cdot \left(1 - \frac{1}{G}\right), \tag{3.2}$$

where P_{IN} is the power of the input signal and G is the gain of the amplifier stage.

The maximum transmitting power level of mobile phones is usually in the region of 1W, and the power level of base stations is a great deal higher. However, modulators or upconversion mixers are only able to generate transmitter signals with a power of below 1 mW. As a result, a substantial power gain is needed in the transmitter chain, which uses cascaded stages to produce the desired output characteristic. The total efficiency of a two-stage cascade is calculated as

$$\eta_{TOT} = \frac{1}{\dfrac{1}{\eta_1 \cdot G_2} + \dfrac{1}{\eta_2}}, \tag{3.3}$$

where η_1 and η_2 are collector/drain efficiencies (not PAEs) of the first and second stages and G_2 is the gain of the last stage. We see from (3.3) that the total efficiency of the system is dominated by the efficiency of the last stage. Let us assume that the efficiency of the last stage is 50% and that it has a gain of 15 dB. If the efficiency of the first stage is now changed from 20% to 30%, the total efficiency increases only by 1% (from 46% to 47%).

The calculations above prove that significant improvements in total efficiency can be obtained by improving the efficiency of the last amplifier stage. Consequently, most effort should be put into the trade-off between efficiency and linearity at this stage. As the first stages can be designed without too much trouble, this book concentrates on the design of the last-stage amplifier. Multistage PA design can also entail some additional difficulties arising from interstage matching, for example, but since these effects are well covered in the literature [1-3], there is no reason to repeat

them in this book, even though the test setups and the analysis presented here can also be extended to multistage amplifiers.

3.2 Linearization

3.2.1 Linearization and Efficiency

Back-off is the traditional way of meeting linearity requirements in class A amplifiers. Once the output power is reduced from its maximum value, both the amount of amplitude conversions and distortion products is reduced. Unfortunately, the back-off reduces efficiency [4], making it an unattractive linearization method for amplifiers. Figure 3.1 presents the level of IM3 as a function of efficiency at various back-off values for a class A amplifier. A third-degree polynomial input-output amplifier model is used in this simulation, and the results indicate that efficiency decreases rapidly when lower IM3 levels are required.

Efficiency and linearity are opposite requirements in traditional power amplifier design, and if the goal is to achieve good linearity with reasonable efficiency, some linearization technique has to be employed. The main idea of linearization is that the power amplifier itself is designed to achieve good efficiency at the expense of linearity, after which the linearity requirements are fulfilled by external linearization. This is demonstrated in Figure 3.1. Let us assume the IM3 specification to be –45 dBc. Without linearization, the amount of back-off that fulfils the IM3 specification would result in an efficiency of approximately 10%. The

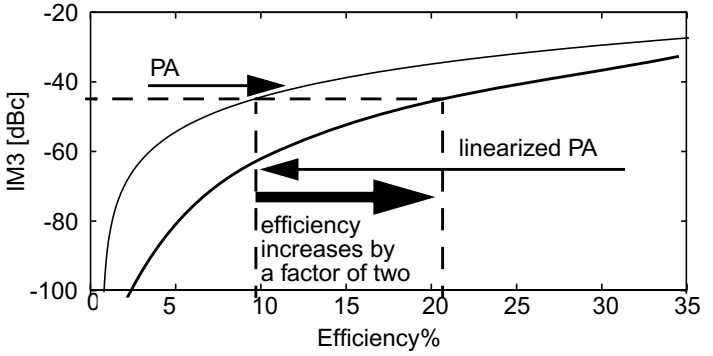


Figure 3.1 Linearity of a PA as a function of efficiency in standalone and linearized configuration.

lower curve presents the linearized IM3 value, with a same linearity achieved at the power amplifier efficiency of better than 20%. In this example, the power consumption of the power amplifier is reduced by more than a factor of two.

The calculation above considers only the power consumption of the amplifier, but in reality linearization also consumes a significant amount of power. Let us assume that the output power is compressed by 0.25 dB, which is a typical value for a modern telecommunications amplifier. Now some $10^{0.25/10} - 1 = 6\%$ additional power is needed in the output to restore the power of the fundamental output signal, and an additional 1% is enough for canceling the approximately –25-dBc IM3 components. The total additional power needed both to restore the fundamental and to cancel the IM3 signals is therefore close to 7% of the output power of the PA, which is not excessive. Unfortunately, it is large enough so that the efficiency and construction of the linearizer circuitry does matter.

3.2.2 Linearization Techniques

Several linearization techniques exist, and they are discussed in more detail in [5-7]. Only the most common categories are briefly explained here, and a short comparison is present in Table 3.1.

Feedback is commonly used, and it can suppress distortion as long as the feedback loop has sufficient incremental gain. To increase the loop gain, baseband error amplifiers in Cartesian [Figure 3.2(a)] or polar form are commonly used. The necessary up- and downconversions inside the loop increase noise sources and loop delays, limiting the stable bandwidth of the linearizer typically to below 100 kHz. As the amplifier is driven close to compression, also the loop gain and bandwidth vary with signal amplitude, complicating the analysis of the system.

Feedforward [Figure 3.2(b)] is commonly used in wideband amplifiers. Here, the distortion generated in the main amplifier is extracted by subtracting the linear contribution from the output of the main amplifier. This distortion signal is amplified by an auxiliary amplifier and finally subtracted from the output. As this arrangement does not contain a feedback loop, it has no stability limitations, but still the bandwidth of the combiners and phase shifters limits the cancellation bandwidth. Unfortunately, the phase shifters and attenuators needed in the feedforward loop are quite tricky to tune automatically, and the use of the linear auxiliary amplifier noticeably reduces the overall efficiency.

Predistortion [Figure 3.2(c)] is based on expanding the signal before the power amplifier, so that the predistorter-amplifier pair appears as a linear circuit. In principle, predistortion is a very power efficient and

wideband linearization method, although it typically needs a slow feedback to adapt the predistorting function. A simple RF predistorter may consist of just a couple of biased diodes, or the predistortion signal can be generated already in the digital baseband using adapted lookup tables.

The last commonly used technique is called either envelope elimination and restoration (EER) or a Kahn transmitter. Here, the amplitude information is removed from the carrier by limiters and then returned by modulating the power supply of the power amplifier. Hence, rail-to-rail driving and high efficiency can be achieved. Unfortunately, EER is very sensitive to any time or phase difference between the carrier path (transmitting phase information only) and the supply modulating path, containing amplitude information only.

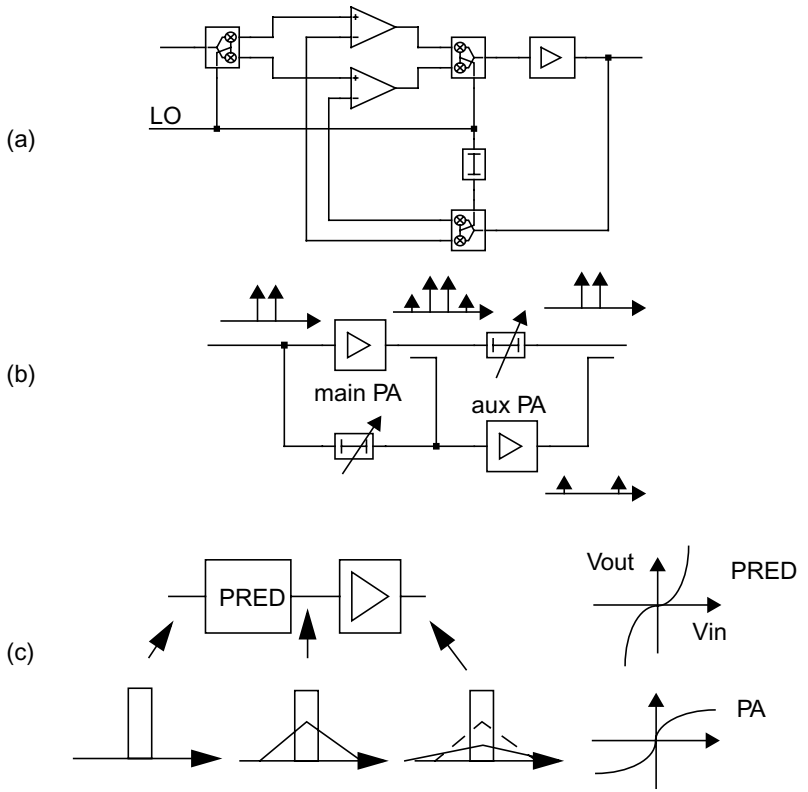


Figure 3.2 Linearization principles: (a) Cartesian feedback, (b) RF feedforward, and (c) predistortion (PRED means a predistorter).

3.2.3 Linearization and Memory Effects

The complexity of different linearization techniques vary, and so does their sensitivity to memory effects. A brief comparison of the presented linearization techniques is presented in Table 3.1.

Table 3.1
Comparison of Different Linearization Techniques

	Complexity	Efficiency	Band-width	Cancell. perform.	Main cause of memory effects
Cartesian feedback	Moderate	High	Narrow	High	Loop bandwidth
Feedforward	High	Moderate	High	High	Passive components
EER	Moderate	High	Moderate	Low	Time delays
RF predistortion	Low	High	High	Low	Power amplifier
Digital predistortion	High	Moderate	Moderate	Moderate	PA & BB and IF filters

Feedback systems like Cartesian feedback are quite insensitive to memory effects in the power amplifier, as they sample the output distortion as it is and try to cancel it with sufficient loop gain. However, to stabilize the loop, the bandwidth must be limited, and this reduces the cancellation far away from the carrier. The feedforward technique also samples directly the output distortion, and then amplifies and subtracts it from the output. Here, the dominant memory effects come from the frequency response of the auxiliary path and passive components, both of which may reduce cancellation far from the carrier. The main concern in feedforward amplifiers is the complexity of tuning the two subtracting/summing loops.

EER technique relies on the matching of two signal paths, one for phase and the other for amplitude information. The main concern tends to be the delays and linearities of these signal paths.

Digital and analog RF predistortion are tempting alternatives in the sense that the distortion is corrected *before* the power amplifier; hence, the output power of the linearizer circuit is smaller and its efficiency is not

such an issue as, for example, in the auxiliary amplifier of a feedforward system. However, predistortion systems rely on exact inverse replication of the nonlinearity of the power amplifier, which means high sensitivity not only to memory effects but to drifting as well. Typically, some form of slow adaptation is needed for the predistorter. Digital predistorter is naturally more flexible, but it requires a high bandwidth and dynamic range from the digital baseband, and also all baseband and intermediate frequency (IF) filters between the predistorter and the power amplifier contribute to the memory effects (see [8]).

Much of the discussion in this book is related to implementing analog RF predistortion systems, or reducing the memory effects of the amplifier to such a low level that a simple memoryless digital predistorter can be used. Still, the analysis techniques presented are applicable to most of the other linearization techniques as well.

A simplistic way to look at any linearizer is to consider it a canceler: a certain amount of distortion is generated and it must be canceled with exactly the opposite phase replica of that distortion. Good cancellation performance places very tight requirements on the amplitude and phase match between the distortion components of the amplifier and the signal components generated in the linearizer. This cancellation is demonstrated in Figure 3.3(a). The power of the residual IM component can be calculated using the cosine rule, and the required matching for a given degree of cancellation is shown in (3.4), where $\Delta\phi$ and ΔA are the phase and amplitude errors, respectively. Figure 3.3(b) shows the corresponding numerical values. To achieve a 25-dB reduction in the IM level, for example, the phase error cannot exceed 2º to 3º and a gain matching $\Delta A/A$ (flatness) better than 0.25 dB (3% error in amplitude) is needed over the entire signal and IM band [9, 10].

$$\mathrm{CANC} = 10 \cdot \log(1 - 2(1 + \Delta A/A)\cos(\Delta\phi) + (1 + \Delta A/A)^2) \quad (3.4)$$

Figure 3.3 clearly illustrates the importance of memory effects. In an amplifier with memory effects, the amplitude and phase of the distortion components vary as functions of modulation frequency (the tone spacing in a two-tone signal) and amplitude. Cancellation signals must track the actual IM signals of the amplifier over the modulation bandwidth, and tracking errors at any modulation amplitude or frequency will cause a deterioration in cancellation performance. This is illustrated in Figure 3.4. The relative phase of the IM signals varies with the modulation frequency, but in simple analog predistorters, for example, predistortion signals are usually memoryless with fixed phase. This problem is often avoided by using a

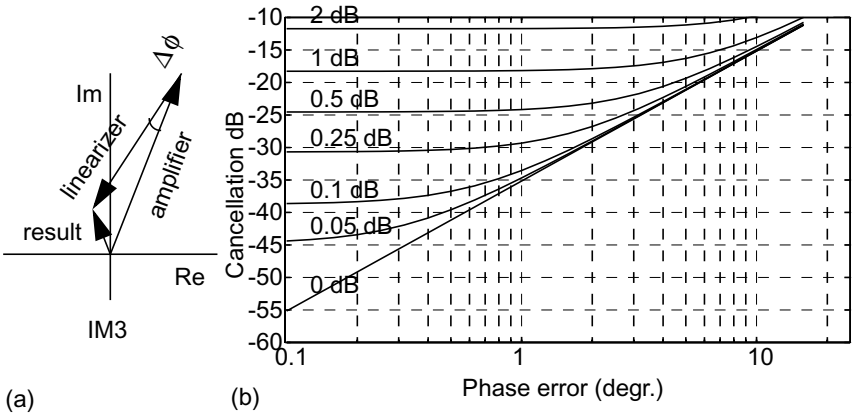


Figure 3.3 (a) Principle of distortion cancellation and (b) the achieved cancellation as a function of phase and amplitude error. From [9].

more complicated digital predistortion algorithm, a feedforward amplifier – or by a power amplifier with a low amount of memory effects.

Distortion components are deterministic signals that vary with the instantaneous amplitude and modulation frequency of the signal. Nevertheless, they always behave similarly under similar conditions. The main contribution of this book is in finding out how distortion components behave under varying signal conditions. This can aid to improve the amplifiers so that good cancellation is achieved using simple RF predistorter type linearization techniques that normally do not provide enough cancellation. By carefully studying their distortion behavior, cancellation can be improved up to 20 to 30 dB that corresponds to the cancellation performance of more sophisticated linearization techniques. If this is achieved, more simple and low-power linearization techniques can be used.

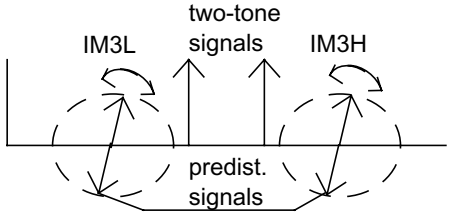


Figure 3.4 Principle of distortion cancellation and its sensitivity to memory effects. © IEEE 2001 [11].

3.3 Electrical Memory Effects

To determine the mechanisms of memory effects, it is important to investigate why the real power amplifier device differs from the polynomial input-output model. This section first presents the impedance definitions of transistor amplifiers. Then, distortion composition is analyzed in more detail and compared with the single polynomial model. Finally, the effects of matching impedances are discussed in terms of memory effects.

The notations for the common emitter/source BJT and MESFET amplifiers given in Figure 3.5 are used throughout this book. Z_B(match) is the driving impedance of the stage, from which the base bias impedance Z_B(bias) is excluded. These two correspond to the impedance that is measured by a network analyzer (NWA) upon disconnecting the transistor. Z_B(int) is the bias-dependent internal base impedance. Similarly, the external collector impedance consists of a load impedance Z_L and a

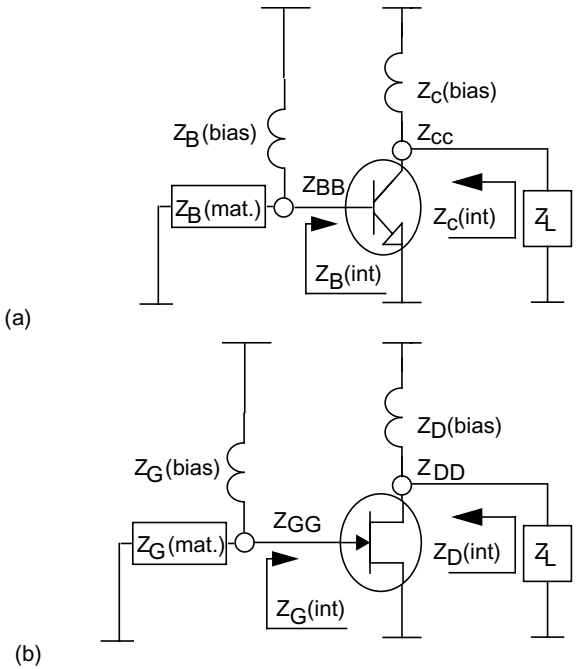


Figure 3.5 Definition of impedances in (a) a CE BJT amplifier and (b) a CS MESFET amplifier. From [12].

collector bias impedance Z_C(bias), which are both measurable from the collector node. Z_C(int) is the internal collector impedance. However, node impedance refers to the impedance level of the node, and the impedance of the base and collector nodes can be calculated by

$$Z_{BB} = Z_B(\text{match}) \parallel Z_B(\text{bias}) \parallel Z_B(\text{int}) \tag{3.5}$$

and

$$Z_{CC} = Z_L \parallel Z_C(bias) \parallel Z_C(int). \tag{3.6}$$

In the same way, changing the names of the terminals allows us to find equations for the node impedances of the MESFET presented in Figure 3.5(b). Instead of source impedance, the term input impedance will be used for both the BJT and the MESFET to make a clearer distinction between it and the source terminal of the MESFET. Equations (3.5) and (3.6) describe the node impedance outside the transistor, although the impedance seen by the internal distortion generator is of primary interest here. These internal impedances will be calculated and used in the simulations and analyses presented in later chapters of this book.

Real power amplifier devices contain more than one nonlinearity mechanism. As a result of their interaction, nonlinear responses are not just output signals as assumed in (2.3); rather, they act as inputs to other nonlinearities and are thus capable of generating new nonlinear responses. Thus, to improve our understanding of distortion mechanisms, the following simplified presentation regards the transistor amplifier as a cascade of two nonlinearities. Although this model lacks the feedback effects of real amplifiers, it is informative and provides an insight into the composition of distortion.

A cascade of two connected Volterra kernels can be presented in the form of a block diagram given in Figure 3.6. Block H describes the base voltage as a function of the input signal, and block F the collector voltage as a function of the base voltage. H_1, H_2, and H_3 correspond to the different order blocks represented by the coefficients a_1, a_2, and a_3 in (2.3), because polynomial input-output models reduce kernels to polynomial coefficients. The generation of IM3 by means of third-degree nonlinearities is straightforward [13]. First, the third-order block H_3 of the base nonlinearity generates an IM3 signal at the base, which is linearly amplified in F_1 and, second, the linear signal at the base goes to the cubic nonlinearity of the transconductance F_3, also producing IM3. The generation of IM3 by cascaded second-degree nonlinearities is somewhat more complicated.

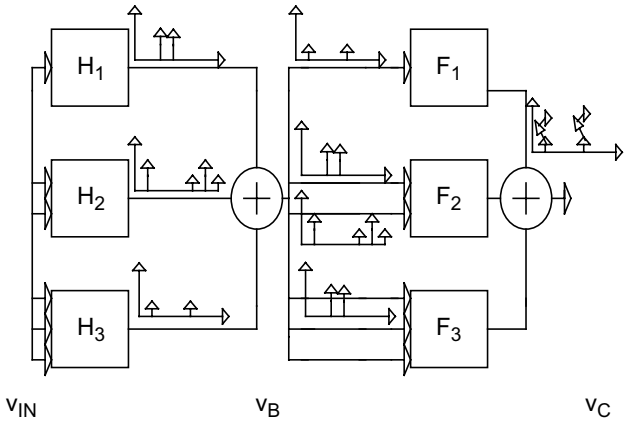


Figure 3.6 Representation of the cascade connection with Volterra operators of the order of one to three. © IEEE 2000 [14].

First, an envelope component is generated at the base in H_2 which, in turn, creates the IM3 component together with the linear signal at the base H_1 in the quadratic nonlinearity of the transconductance F_2. Similarly, mixing from the second harmonic adds to IM3.

A frequency domain combination of the different order spectrums is given in Figure 3.7. Figure 3.7(a) presents the output of the first block that includes the same frequency components as the third-degree polynomial model (2.3). The amplitude of the spectral components can be found from Table 2.4. This multitone signal is the input signal for the second block, and the output IM3 now combines with other frequency components. The envelope signal ω_2–ω_1 and the upper two-tone signal ω_2, for example, will be mixed in the quadratic nonlinearity of the latter block, which results in the generation of the upper IM3 signal (and ω_1 compression, as well). Similarly, the second harmonic of the upper input signal $2\omega_2$ and the lower input signal from the negative frequency side $-\omega_1$ will also mix to the IM3 signal. As a result, IM3 sidebands are affected not only by the fundamental voltage waveforms, but also by the voltage waveforms of the different nodes at the envelope and second harmonic frequencies ω_2–ω_1 and $2\omega_2$.

The question is how to control the voltage waveforms of the different nodes and frequency components. Since the nonlinearities of the circuit components can be regarded as current sources, as explained in Chapter 2, their voltage waveforms can be affected by node impedances. The composition of IM3 in the real power amplifier device is sketched in Figure

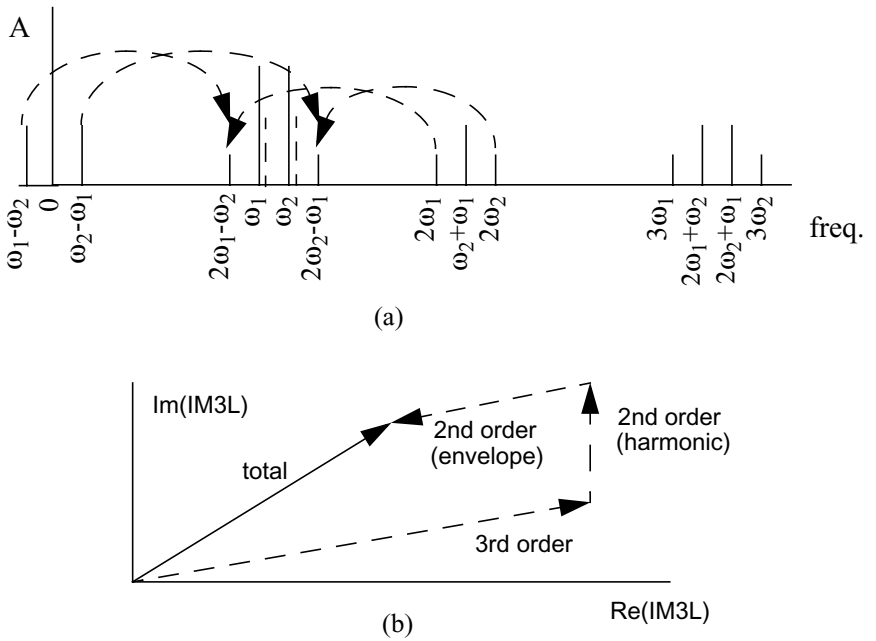


Figure 3.7 (a) Spectral components produced by nonlinearities up to the third order and (b) composition of IM3. © IEEE 2001 [11].

3.7(b) with nonlinearities up to the third degree [15]. IM3 is largely generated by cubic nonlinearities, which are affected by fundamental impedances and signal levels. However, quadratic mechanisms that mix the envelope and second harmonic frequencies with the fundamental tones also have a significant effect on IM3 distortion, and these can be controlled by node impedances at these out-of-band frequencies.

Electrical memory effects are caused by frequency-dependent envelope, fundamental or second harmonic node impedances. Figure 3.8 gives the measured gate node impedances for the MESFET amplifier in the dc, fundamental, and second harmonic bands. The center and maximum modulation frequencies are 1.8 GHz and 20 MHz, respectively, which means that the dc band is important up to 20 MHz or beyond. The interesting fundamental band is between 1.77 GHz and 1.83 GHz, because the entire IM3 band of 60 MHz is relevant in terms of IM3 distortion. The second harmonic band lies between 3.58 GHz and 3.62 GHz. The fundamental impedance can easily be kept constant over the entire

modulation frequency range, because it is just 0.3% of the center frequency in our example. Also, the second harmonic band is quite narrow, and impedance matching is simple, provided that no harmonic traps are used. Such traps cause tremendous impedance variations and may cause significant memory effects. As the fundamental and second harmonic impedances play a minor role, memory effects are for the most part produced by envelope impedances. The envelope frequency varies from dc to 20 MHz, and the gate node impedance, for example, must be constant or very low over this region to eliminate memory effects. This is not the case in the practical implementation presented in Figure 3.8, however, where the gate impedance at the envelope frequency varies by approximately two decades.

There is one important difference between the dc and the other frequency bands. If the center frequency of the system changes, both the fundamental and the second harmonic impedance change, while the envelope impedance remains the same. In other words, if a significant amount of memory is produced by the fundamental or second harmonic bands, the shape of these memory effects will change with the frequency channel. Despite that, it can be concluded that with careful design, memory effects introduced by various terminal impedances can be limited to those converted from the envelope frequency. A thorough analysis of distortion mechanisms will be presented in Chapter 4, where the distortion mechanisms and memory effects of the BJT and MESFET amplifiers are analyzed in detail.

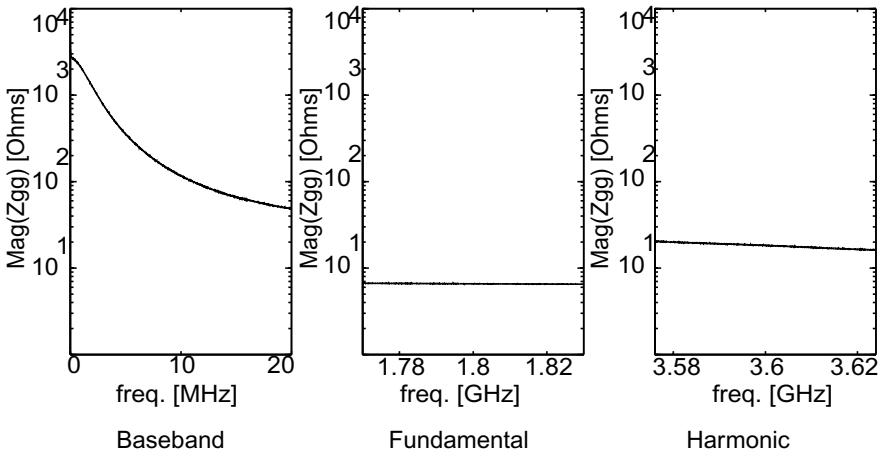


Figure 3.8 Measured magnitude of the Z_{GG} of the MESFET amplifier. From [12].

3.4 Electrothermal Memory Effects

Electrothermal memory effects are caused by electrothermal couplings, which affect low modulation frequencies up to the megahertz range. The dissipated power of the BJT can be expressed as

$$p_{\mathrm{DISS}}(t) = v_{\mathrm{CE}}(t) \cdot i_{\mathrm{CE}}(t), \tag{3.7}$$

where v_{CE} is the collector-emitter voltage and i_{CE} the collector-emitter current. Since two first-order fundamental signals are multiplied together, the spectrum of the dissipated power always includes second-order signal components (i.e., dc, envelope, sum, and second harmonics [16]). The temperature variations caused by the dissipated power are determined by the thermal impedance (Z_{TH}), which describes the ratio between temperature rise and heat flow from the device. Due to the nonzero mass of the component, thermal impedance in the active device is not purely resistive, but forms a distributed lowpass filter with a wide range of time constants.

This means that the temperature changes caused by the dissipated power do not occur instantaneously, but due to the mass of the semiconductor and the package, a frequency-dependent phase shift always exists. Moreover, the surface of the silicon reacts surprisingly quickly, and thermal effects can be obtained in bandwidths up to 100 kHz to 1 MHz [16-20]. Furthermore, since heat in the chip flows mostly vertically [21], it can be assumed that self-heating within the component produces more memory effects than the heat generated by surrounding heat sources.

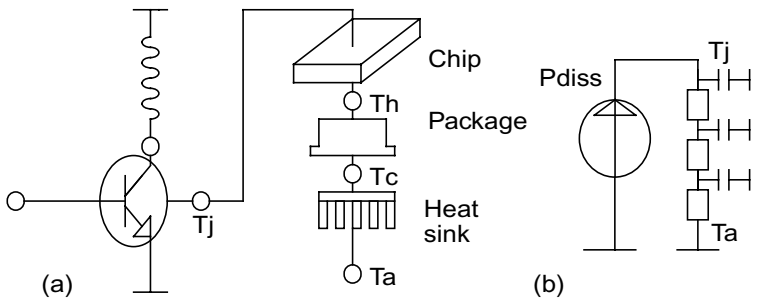


Figure 3.9 Heat flow from the device: (a) physical and (b) electrical lumped element models. © IEEE 2001 [11].

A finite element model is employed here to simulate the thermal impedances of the package presented in Figure 3.9. For the sake of simplicity, the structure of the package presented in Figure 3.10 is modeled using brick elements. The silicon chip measures 600 μm by 600 μm by 300 μm and the emitter of the transistor 400 μm by 400 μm. The thicknesses of the copper lead frame and the moulding compound are 100 μm and 1000 μm, respectively, and the temperature of the bottom of the moulding compound is assumed to be constant. The structure includes 1859 nodes, that is, the grid is 50 μm in length and width and the thickness of the grid of the chip, the lead frame, and the molding compound are 50 μm, 50 μm, and 500 μm, respectively.

Figure 3.11 illustrates three simulated thermal impedances for the entire structure. The first one is simulated at the center of the active area (marked with a square), the second at the corner of the chip (diamond), and the third a fitted lumped model with three time constants (cross). The figure indicates that the surface of silicon reacts quickly, and a thermal impedance

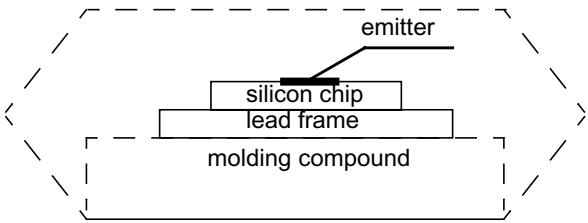


Figure 3.10 Simplified package structure. From [22].

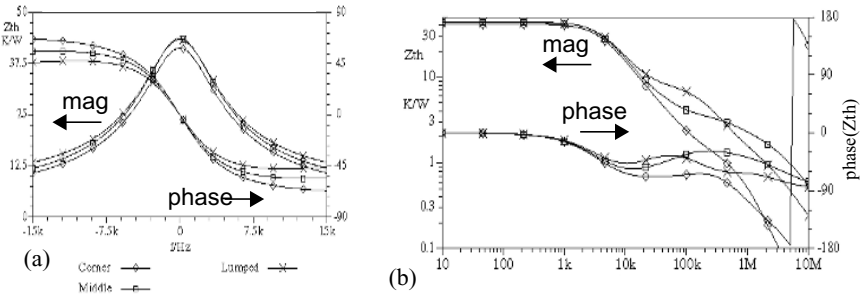


Figure 3.11 Simulated thermal impedance at different locations of the integrated circuit (IC): (a) on a two-sided linear frequency axis, and (b) on logarithmic frequency axis. From [22].

of several ohms (approximately 10% of R_{th}) can be seen up to frequencies of 1 MHz. To illustrate the odd phase response, Figure 3.11(a) is plotted on two-sided frequency axis, while in Figure 3.11(b), typical log-log magnitude and log-lin phase plots are shown. A more detailed analysis of this thermal impedance can be found in [23-25]. Thus it suffices to say here that the effects of the package and the heat sink are important from the point of view of thermal resistance, which determines the average temperature rise caused by self-heating. From the ac behavior point of view, layers near the active area (silicon and lead frame) are more dominant, because the responses of package and heatsink are usually far too slow compared to microsecond range variations in power dissipation. Thermal impedance simulations for the GaAs MESFET are presented in [26, 27], where similar time constants are reported for silicon to those that were found in the simulations presented here.

Since power is dissipated at dc, fundamental, and second-order signal frequencies, but only dc and envelope components of the dissipated power fit into the passband of the thermal filter, the temperature of the chip takes the following simple form

$$T = T_{AMB} + R_{TH} \cdot p_{DISS}(0Hz) + Z_{TH}(\omega_1 - \omega_2) \cdot p_{DISS}(\omega_1 - \omega_2). \quad (3.8)$$

The temperature of the chip consists of three components: one is simply the ambient temperature T_{AMB} and the other two consist of the thermal resistance multiplied by the dc power dissipation, and the envelope component multiplied by the thermal impedance at that frequency. It is interesting to note that the third term in (3.8) includes frequency, which means that the temperature variations at the surface of the chip also depend on the bandwidth of the signal. If any of the electrical parameters of the transistor are affected by temperature, thermal memory effects are unavoidable. This mechanism in which dynamic self-heating causes electrical distortion is known as thermal power feedback (TPF) [28].

A block diagram of TPF is shown in Figure 3.12, in which the basic amplifier is considered to be a polynomial input-output stage. Thermal impedance describes the relationship between dissipated power and temperature, and block K describes the relationship between temperature and the gain of the amplifier. Only the gain of the amplifier is considered to be temperature-dependent in this behavioral model. In practice, however, the output conductance [29] and the capacitances are also temperature-dependent at the transistor level, as will be seen in Chapter 4. Since some of the circuit parameters of the transistor are always functions of temperature, TPF cannot be avoided. TPF is a very difficult problem to

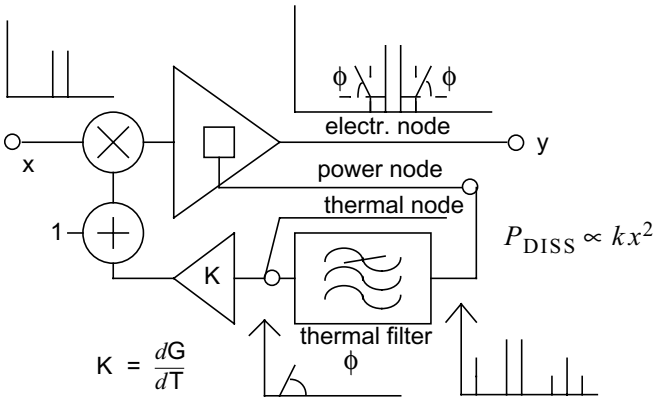


Figure 3.12 Block diagram of thermal power feedback (TPF). © IEEE 2001 [11].

compensate for, because the exact chip temperature cannot usually be measured. For example, temperature-compensated external bias networks are incapable of detecting junction temperatures, and hence far too slow to compensate for changes therein and cannot offer an improvement in thermally induced distortion.

We now demonstrate thermal power feedback. We choose 0.15-*j*0.15 as the normalized cubic nonlinearity coefficient a_3/a_1, a value which corresponds to an IM3 level of –40 dBc at the signal level employed. The dG/dT is –0.6%/K and the thermal impedance is taken from Figure 3.11. A two-tone input signal is modulated by thermally induced gain variations at the envelope frequency, thereby generating IM3 sidebands. Since the phase response of the thermal filter at the positive envelope (generating IM3H) is opposite to that at the negative envelope (IM3L), the IM3 sidebands produced by TPF turn in opposite directions as a function of modulation frequency, as shown in Figure 3.13. A linearity decrease of several decibels is observed at low modulation frequencies, and a phase imbalance exists between the sidebands at some relatively low values.

3.5 Amplitude Domain Effects

Memory effects that are introduced into power amplifiers affect their distortion performance related to both modulation frequency and amplitude. Modulation frequency-dependent effects were used in the previous sections to demonstrate the mechanisms of memory effects, while

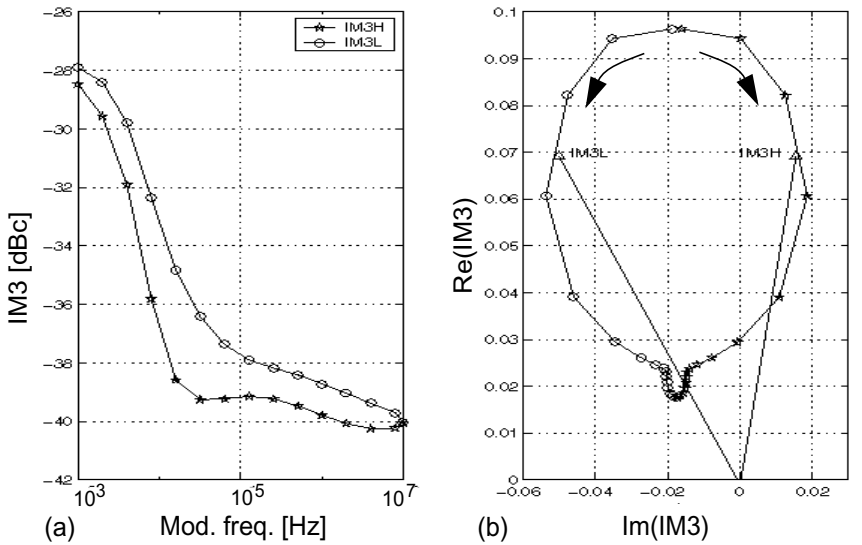


Figure 3.13 IM3 caused by the basic amplifier and TPF. (a) Represents the magnitude of IM3 as a function of modulation frequency and (b) presents the IM3 in real-imaginary coordinates. From [22].

this section examines the effects of amplitude on memory effects. It is important to note that, due to the definition, the memory effects considered here are actually both modulation amplitude and frequency dependent. The term "amplitude dependent" is justified, because effects higher than third-order are taken into account here, and the amount of fifth-order distortion in the IM3 tone depends on signal amplitude. These effects are slightly more complicated than frequency domain memory effects, wherefore third- and fifth-order distortion composition will be studied first without memory effects. They will be taken into account later by considering the amplifier as a cascade of two polynomials with a bandwidth-limited connection. This gives a useful insight into distortion mechanisms for the further simulations to be presented in Chapter 6.

3.5.1 Fifth-Order Analysis Without Memory Effects

Polynomial input-output relations up to the fifth degree can be written as

$$y = a_1 \cdot x + a_2 \cdot x^2 + a_3 \cdot x^3 + a_4 \cdot x^4 + a_5 \cdot x^5, \tag{3.9}$$

where a_1-a_5 are real valued coefficients. By applying a two-tone signal (2.7), the in-band intermodulation products shown in Figure 3.14(a) can be obtained.

The relation between the degree of the nonlinearity (third, fifth,...) and the frequency of the tone (such as IM3, IM5,...) is demonstrated in Figure 3.14(a). The IM5 tones are not affected by third-degree nonlinearities, but IM3 tones are functions of both third- and fifth-degree nonlinearities. This means that at low signal amplitudes, where the fifth-order distortion products can be neglected, the amplitudes of the IM3 tones are proportional to the third power of the input amplitude. With a fairly large signal amplitude, however, fifth-order products (which are dependent on a power of five) will start to affect the IM3 responses. As a result, the 3:1 amplitude estimate will no longer hold, as demonstrated in Figure 3.14(b). If the phases of the third- and fifth-degree coefficients are equal, the fifth-degree nonlinearity will expand the IM3 responses. However, if the phases are the opposite, the IM3 distortion will be locally reduced, as shown in the figure. This explains why notches in the IM3 sidebands have been reported at certain amplitudes [30, 31]. It is also interesting to note that the amplitude of IM3 (25/8) caused by the fifth-degree nonlinearity is five times greater than that of IM5 (5/8). This information is necessary for the identification of amplitude domain memory effects.

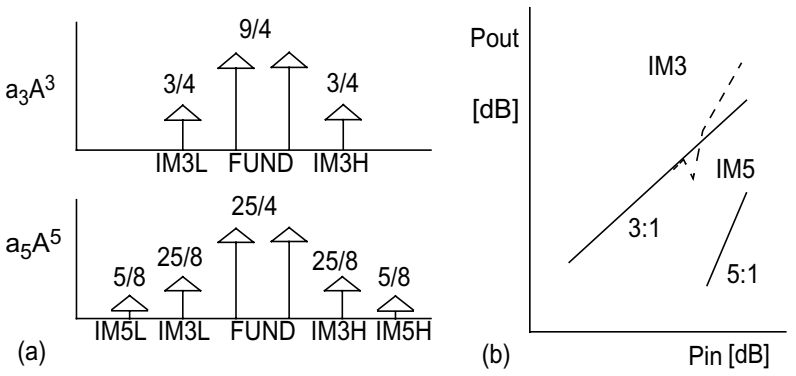


Figure 3.14 (a) Distortion components caused by third- and fifth-degree nonlinearities and (b) amplitude of IM3 and IM5 components as function of input amplitude. From [32].

3.5.2 Fifth-Order Analysis with Memory Effects

The cascade representation of nonlinear systems introduced in Section 3.3 will now be extended to fifth-degree nonlinearities. As noted in the previous section, IM3 is also generated by fifth-degree nonlinearities. The connection between the two blocks is bandwidth-limited to provide an insight into amplitude-dependent memory effects. A block diagram is given in Figure 3.15(a).

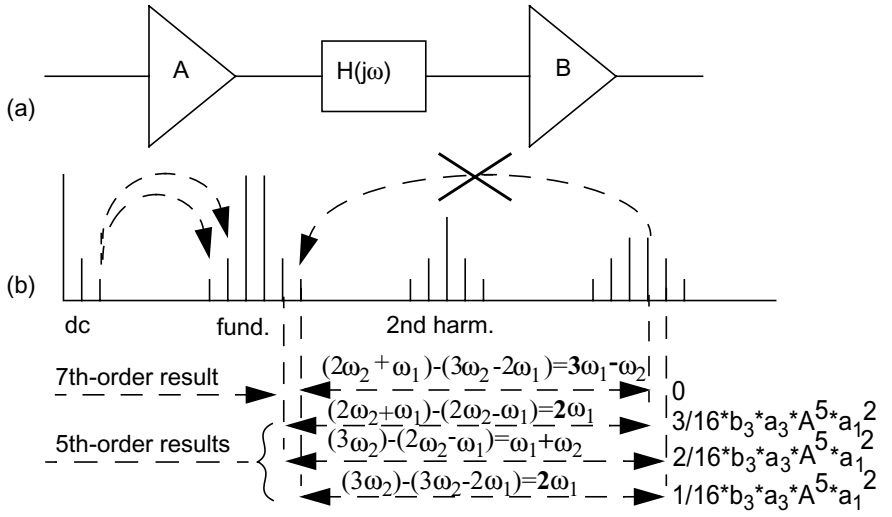


Figure 3.15 (a) Cascade of two polynomials, and (b) mixing of distortion of block A in block B. From [32].

The filter $H(j\omega)$ will be ignored at first, and since a cascade of two polynomials is still just a polynomial; no difference exists between it and the single polynomial presented in Section 3.5.1. The coefficients are different, but the 5:1 ratio between the fifth-order IM3 and IM5 terms remains constant. Four frequency bands have an impact on the IM responses: dc, fundamental, second harmonic, and third harmonic. Signals occur at the fourth and fifth harmonics as well, but since the distortion analysis is truncated at the fifth-order, these signal components do not affect the in-band intermodulation distortion. Each band consists of three or more individual tones. Figure 3.15(b) presents the significant signal components of the IM3 and IM5 distortion.

Next, the effects of the third harmonic to the in-band intermodulation distortion will be studied. The third harmonic band consists of frequency components at six frequencies, of which the outermost ($4\omega_1$-ω_2 and $4\omega_2$-ω_1) are fifth or higher order distortion products. This is obvious, because these frequencies cannot be combined from three fundamental (ω_1 or ω_2) frequencies only. However, the middle four spectral components consist of both third- and fifth-order terms, of which the third-order ones are written in Table 2.5. We now neglect all fifth-order effects at the third harmonic, because mixing them to the fundamental band corresponds to at least a seventh-order effect. At this stage, we have to mix down to the intermodulation band the spectral components at $3\omega_1$, $2\omega_1+\omega_2$, $2\omega_2+\omega_1$, and $3\omega_2$ with amplitudes of $a_3/4$, $3a_3/4$, $3a_3/4$, and $a_3/4$. Furthermore, it is assumed that the role of the second harmonic in the middle of the blocks is negligible (i.e., the second and third harmonics do not mix back to the intermodulation band in the latter block B). The intermodulation distortion at the output from the third harmonic is now composed of the following phenomena: the third harmonic band generated by a_3 and the fundamental components (taken two times) produced by a_1, mixed in the cubic nonlinearity of the latter block B. This is presented in Figure 3.16.

This particular distortion mechanism produces the IM3 and IM5 components of $5/16*a_1^2*a_3*b_3*A^5$ and $1/16*a_1^2*a_3*b_3*A^5$ that contribute the IM3 and IM5 in the same amplitude ratio of five. In other words, if the third harmonic is filtered out with filter H, the amount of fifth-order IM3

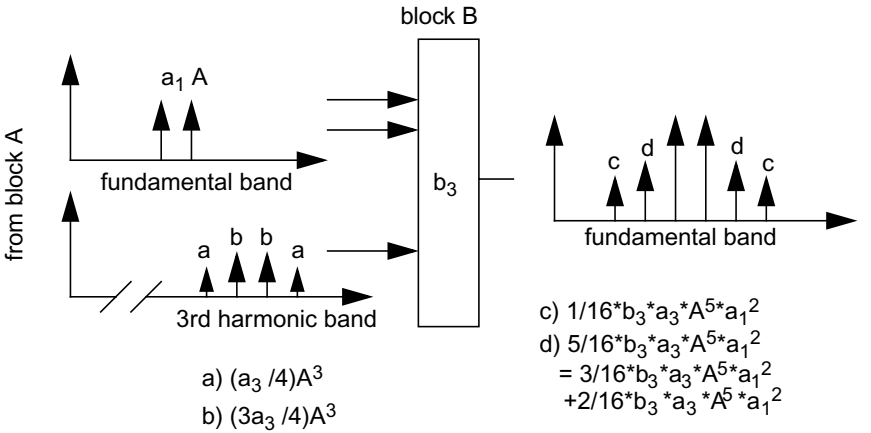


Figure 3.16 The mechanism of IM3 and IM5 downconverting from the third harmonic band.

and IM5 changes, but the ratio of five obtained from the memoryless distortion composition remains constant. This observation can be generalized as follows: the same 5:1 ratio is also true for other distortion mechanisms that are converted from the dc and the second harmonic bands. Once the frequency band is filtered out, the ratio remains the same. Moreover, when the filtering inside the frequency band is flat (i.e., all spectral components are attenuated identically), the 5:1 ratio exists between the IM3 and IM5 responses caused by the fifth-degree nonlinearity. It is still important to emphasize that the ratio between the total IM3 and IM5 changes, because IM3 also comprises a third-order component that is not a function of third harmonic filtering.

Let us now take a look at filtering inside the frequency band. Figure 3.15 sketches the important signal components at the third harmonic band in terms of intermodulation distortion. The upper third harmonic $3\omega_2$ mixes down to IM3 and IM5 at the amplitudes of 2/16 and 1/16, but the sum frequency $2\omega_2+\omega_1$ only mixes to IM3 at the amplitude of 3/16. There is no mixing down to IM5, because the difference between the $2\omega_2+\omega_1$ and the $3\omega_2-2\omega_1$ tones is $3\omega_1-\omega_2$. Since the $2\omega_1+\omega_1$ and $3\omega_1-\omega_2$ are at least third- and fourth-order effects, IM5 converting down from the $2\omega_1+\omega_1$ is at least a seventh-order effect As the spectral components at the third harmonic band are divided into its contributors, it becomes clear that if the sum frequencies $2\omega_1+\omega_2$ and $2\omega_2+\omega_1$ exist, but the harmonics ($3\omega_1$ and $3\omega_2$) are filtered, the value of the ratio between IM3 and IM5 caused by fifth-degree nonlinearities at the third harmonic band is infinity instead of five.

Amplitude-dependent memory effects can be understood as deviations from the 5:1 amplitude ratio for fifth-order IM3 and IM5 responses or as phase differences between the two. This occurs whenever one of the frequency bands is filtered unevenly. To study these effects, the filter $H(j\omega)$ is considered to be tilted around the dc band, because it is the most important source of memory effects in practical applications. The composition of fifth-order distortion mechanisms to IM3 and IM5 is studied as a function of tone spacing. Since the frequency of the fourth-order envelope $2\omega_1-2\omega_2$ is twice that of the second-order envelope $\omega_1-\omega_2$, the IM3 and IM5 generated by the fourth-order envelope will include a significant amount of memory that is a function of signal amplitude.

This means that the 5:1 ratio fails to hold whenever the filter is tilted, and if the phase of the filter is not equal at the second and fourth-order envelopes, a phase imbalance will exist between the fifth-order contributors, as seen in Figure 3.17(a). A similar conclusion can be reached concerning the other frequency bands. In other words, the spectral components that are furthest away from the center of the frequency band are the most harmful in terms of amplitude-dependent memory effects.

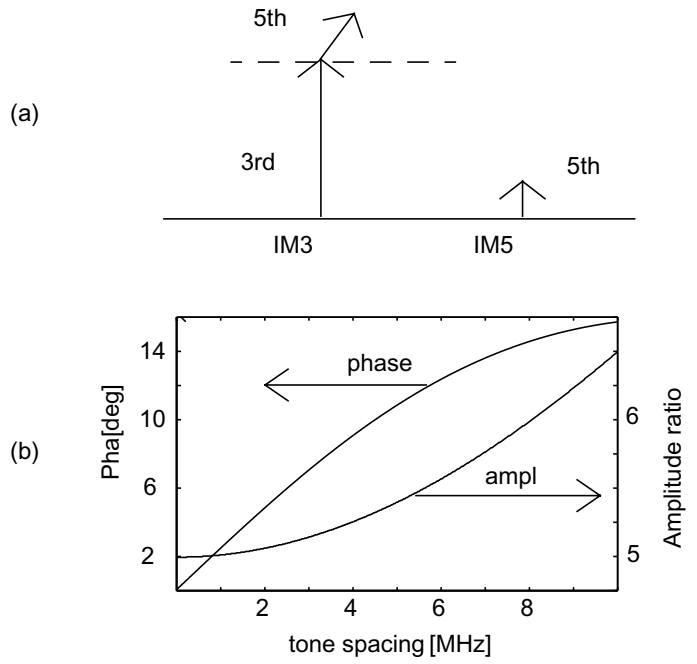


Figure 3.17 (a) Composition of fifth-order distortion and (b) memory effects caused by it. From [32].

The ratio between the fifth-order IM3 and IM5 contributors as a function of modulation frequency is presented in Figure 3.17(b). If the modulation frequency is close to zero, no memory effects occur and the amplitude ratio is 5:1. The ratio starts to increase with increasing modulation frequency, however, and acquires the value of 6 at 8 MHz. Also, a phase difference of 16º exists between the fifth-order IM3 and IM5 contributors at high modulation frequencies. It is important to note that these values are closely dependent on the nonlinearity coefficients and on the filtering and that the result presented here cannot be generalized. It is evident, however, that deviations from the memoryless approximation are unavoidable, if the maximum modulation frequency is in the MHz range.

One important feature of amplitude-dependent memory effects arises from the third harmonic band. Let us assume that all frequency bands up to the second harmonic are flat over the whole range of distortion bands, so that no memory effects exist at moderate amplitude levels, because the higher than third-degree nonlinearities are practically zero. If the signal

amplitude is now increased, fifth-order mechanisms will start to have an effect, and signal components from the third harmonic band will be converted down to IM3. This means that if only the third harmonic band is tilted, memory effects will occur exclusively at high signal levels, where fifth-order distortion begins to play a significant role.

3.6 Summary

Once the actual transfer function of the amplifier is affected by the bandwidth of the applied signal, it will exhibit memory effects. There are some ways of measuring these effects, and in this book intermodulation signals as a function of tone difference of a two-tone signal are measured. If the amplitude and/or phase of the IM signals is affected by the tone difference, the amplifier exhibits memory effects.

Smooth memory effects are not usually harmful to the linearity of the PA itself. A phase rotation of 10º to 20º or an amplitude change of less than 0.5 dB as a function of modulation frequency has no dramatic effect on the linearity of the device, but the situation is completely different when predistortion type linearization is employed to cancel out the IM sidebands. This undertaking requires an extremely accurate amplitude and phase matching between the distortion components and the cancelling signals. If the IM3 components rotate as a function of modulation frequency, for example, but the canceling signals do not, the cancellation performance of the linearization method may be inadequate for wideband signals. Different linearization techniques have different sensitivities to memory effects, and feedback or feedforward type linearizers, for example, are not very sensitive to amplifier behavior. However, RF and digital baseband linearizers have potential for both high efficiency and wide bandwidth – provided that the deterioration in the cancellation performance due to memory effects can be cured.

This chapter introduced two memory effects: electrical and thermal. Electrical memory effects are produced by nonconstant node impedances within frequency bands. Most of these effects are generated by a frequency-dependent envelope impedance, caused usually by the bias impedance. Thermal memory effects, for their part, are generated by the junction temperature, which is modulated by the applied signal. Since the chip temperature varies at the envelope frequency and some of the electrical parameters of the transistor are affected by the temperature, gain modulation and IM3 components are generated. These memory effects occur, because the temperature rise caused by dissipated power is highly

dependent on the modulation frequency, which also determines the behavior of the IM3 components.

If the signal amplitude is reasonably large, fifth- and higher order distortion mechanisms will have an impact on IM3 performance. Fifth-order distortion will affect both the IM3 and IM5 sidebands, and in the polynomial case, the fifth-order IM3 and IM5 contributions will have 5:1 amplitude ratio and a phase difference of zero. Amplitude-dependent memory effects that change this situation, however, arise from two phenomena: first, the frequency bands become wider, because fifth-order signal components are wider than third-order components alone and, second, also the third harmonic band converts down to IM at high signal levels. These effects are called amplitude-dependent memory effects, because the amount of fifth-degree nonlinearity in the IM3 components is dependent on the amplitude, and the total extent of amplitude-dependent memory effects is the sum of the memory effects converted from all frequency bands.

Memory effects can be visualized by a vector presentation of IM3. Presenting the contributors to IM3, such as the envelope, fundamental, and second harmonic components, as individual vectors instead of the total result, makes it possible for us to identify the causes of the memory effects. This book will make extensive use of this presentation format.

3.7 Key Points to Remember

1. Efficiency and linearity are opposite design goals in traditional RF power amplifier design.
2. The idea of linearization is that the power amplifier itself is designed to be not linear enough to achieve good efficiency, after which the linearity requirements are fulfilled by external linearization.
3. The implementation and complexity of the linearizer affect the overall efficiency of the transmitter.
4. The cancellation performance of especially the simple predistortion type linearization techniques is sensitive to memory effects generated in the RF power amplifier.
5. There are two types of memory effects: electrical and electrothermal.
6. Distortion appears as currents, and the terminal voltages caused by distortion can be shaped by the terminal impedances.

7. Lower order distortion may mix up to higher order distortion (i.e., *N*th order distortion can be generated by lower degree nonlinearities).
8. Electrical memory effects are caused by nonconstant terminal impedances at dc, fundamental, and harmonic bands, of which those within the dc band are the most harmful, because bias impedances are strongly frequency dependent.
9. The thermal impedance describes the temperature increase caused by dissipated power, and it has a wide range of time-constants and a significant magnitude up to the megahertz frequency range in RF power transistors.
10. Electrothermal memory effects are caused by dynamic temperature variations at the top of the chip that modifies the electrical properties of the transistor at the envelope frequency (tone difference frequency). As a result, IM3 signals that depend on thermal impedance are generated.
11. Fifth-order distortion affects IM3 components, and the memory effects of fifth-order distortion can also be seen in IM3 components.
12. Since fifth-order effects on IM3 are an amplitude-dependent phenomenon, memory effects also become amplitude dependent.
13. With higher-order distortion, the frequency bands get wider and new harmonic bands mix down to fundamental. This causes the memory effects to vary with signal level.

References

[1] Mori, K., et al., "An L-band high-efficiency and low-distortion power amplifier using HPF/LPF combined interstage matching circuit," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 48, No. 12, 2000, pp. 2560-2566.

[2] Cripps, S., *RF Power Amplifiers for Wireless Communications*, Norwood, MA: Artech House, 1999.

[3] Gonzalez, G., *Microwave Transistor Amplifiers: Analysis and Design*, Englewood Cliffs, NJ: Prentice-Hall, 1997.

[4] Krauss, H., *Solid State Radio Engineering,* New York, NY, John Wiley & Sons, 1980.

[5] Kenington, P.B., *High Linearity RF Amplifier Design,* Norwood, MA: Artech House, 2000.

[6] Raab, F., et al., "Power amplifiers and transmitters for RF and microwave," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 50, No. 3, 2002, pp. 814-826.

[7] Cripps, S., *Advanced Techniques in RF Power Amplifier Design*, Norwood, MA: Artech House, 2002.

[8] Weiyun S., L. Sundström, and B. Shi, "Spectral sensitivity of predistortion linearizer architectures to filter ripple," *Proc. IEEE 2001 Vehicular Technology Conference*, Vol. 3, pp. 1570-1574.

[9] Rahkonen, T., and J. Vuolevi, "Memory effects in analog predistorting linearizing systems," *Proc. Norchip 1999 Conference*, Oslo, Norway, November 8-9, 1999, pp. 114-119.

[10] Morris, K.A., and J.P. McGeehan, "Gain and phase matching requirements of cubic predistortion systems," *IEE Electronics Letters*, Vol. 36, No. 21, 2000, pp. 1822-1824.

[11] Vuolevi, J., T. Rahkonen, and J. Manninen, "Measurement technique for characterizing memory effects in RF power amplifiers," *IEEE Trans. on Microwave Theory and Measurements*, Vol. 49, No. 8, 2001, pp. 1383-1389.

[12] Vuolevi, J., "Analysis, measurement and cancellation of the bandwidth and amplitude dependence of intermodulation distortion in RF power amplifiers," Doctoral thesis, University of Oulu, Oulu, Finland, 2001.

[13] Maas, S., "Third-order intermodulation distortion in cascaded stages," *IEEE Microwave and Guided Wave Letters*, Vol. 5, No. 6, 1995, pp. 189-191.

[14] Vuolevi, J., and T. Rahkonen, "The effects of source impedance on the linearity of BJT common-emitter amplifiers," *Proc. 2000 IEEE International Symposium on Circuits and Systems*, Geneva, Switzerland, May 28-31, 2000, pp. IV-197-IV-200.

[15] Sevic, J., K. Burger, and M. Steer, "A novel envelope-termination load-pull method for ACPR optimization of RF/microwave power amplifiers," *1998 IEEE MTT-S International Microwave Symposium Digest*, Vol. 2, pp. 723-726.

[16] Schurack, E., et al., "Analysis and measurement of nonlinear effects in power amplifiers caused by thermal power feedback," *Proc. 1992 IEEE International Symposium on Circuits and Systems*, Vol. 2, pp. 758-761.

[17] Zhu, Y., et al., "Analytical model for electrical and thermal transients of self-heating semiconductor devices," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 46, No. 12, 1998, pp. 2258-2263.

[18] Hopkins, T., and R. Tiziani, "Transient thermal impedance considerations in power semiconductor applications," *Automotive Power Electronics*, 1989, pp. 89-97.

[19] Zweidinger, D., S. Lee, and R. Fox, "Compact modeling of BJT self-heating in SPICE," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 12, No. 9, 1993, pp. 1368-1375.

[20] Le Gallou, N., et al., "Analysis of low frequency memory and influence on solid state HPA intermodulation characteristics," *Proc. 2001 IEEE International Microwave Symposium*, Phoenix, AZ.

[21] Altet, J. et al., "Thermal coupling in integrated circuits: application to thermal testing," *IEEE Journal of Solid-State Circuits*, Vol. 36, No. 1, 2001, pp. 81-91.

[22] Vuolevi, J., and T. Rahkonen, "Third-order intermodulation distortion caused by thermal power feedback," *Proc. Norchip'99 Seminar*, Oslo, Norway, November 8-9, 1999, pp. 120-125.

[23] Perugupalli, P., Y. Xu, and K. Shenai, "Measurements of thermal and packaging limitations in LDMOSFETs for RFIC application," *1998 IEEE Instrumentation and Measurement Technology Conference*, Vol. 1, pp. 160-164.

[24] Hefner, A., and D. Blackburn, "Simulating the dynamic electrothermal behavior of power electronics circuits and systems," *IEEE Trans. on Power Electronics*, Vol. 8, No. 4, 1993, pp. 376-385.

[25] Hefner, A., "A dynamic electro-thermal model for the IGBT," *IEEE Trans. on Industry Applications*, Vol. 30, No. 2, 1994, pp. 394-405.

[26] Veijola, T., M. Andesson, and A. Kallio, "Parameter extraction procedure for an electrothermal transistor model," *Proc. BEC'96*, Tallinn, Estonia, pp. 71-72.

[27] Veijola, T., and M. Andesson, "Combined electrical and thermal parameter extraction for transistor model," *1997 European Conference on Circuit Theory and Design*, Budapest, Hungary, pp. 754-759.

[28] Lee, S., and D. Allstot, "Electrothermal simulation of integrated circuits," *IEEE Journal of Solid-State Circuits*, Vol. 28, No. 12, 1993, pp. 1283-1293.

[29] Fox, R., S. Lee, and D. Zweidinger, "The effects of BJT self-heating on circuit behavior," *IEEE Journal of Solid-State Circuits*, Vol. 28, No. 6, 1993, pp. 678-685.

[30] Gutierrez, H., K. Gard, and M. Steer, "Nonlinear gain compression in microwave amplifiers using generalized power-series analysis and transformation of input statistics," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 48, No. 10, 2000, pp. 1774-1777.

[31] Hau, G., T. Nishimura, and N. Iwata, "Distortion analysis of a power heterojunction FET under low quiescent drain current for 3.5 V wide-band CDMA cellular phones," *1999 IEEE MTT-S Symposium on Technologies for Wireless Applications*, pp. 37-40.

[32] Vuolevi, J., and T. Rahkonen, "Analysis of amplitude dependent memory effects in RF power amplifiers," *Proc. European Conference on Circuit Theory and Design* (*ECCTD'01*), Helsinki, Finland, August 28-31, 2001, pp. II-37-II-40.

Chapter 4

The Volterra Model

In this chapter, we will first review the most common nonlinear modeling techniques. Section 4.1.2 discusses the properties of the polynomial Volterra modeling technique, and Section 4.2 explains in detail how the purely electrical and electrothermal nonlinearities are written and what terms are needed in the series expansion that models the nonlinearity. Section 4.3 illustrates how distortion of a common emitter amplifier is calculated using the Volterra analysis technique, and as a first case study, Section 4.4 presents the tear-down analysis of all the terms affecting the IM3 distortion in a BJT CE amplifier. Similar analysis is performed to a MESFET amplifier in Section 4.5.

4.1 Nonlinear Modeling

To be able to analyze the nonlinear behavior of an amplifier, we need accurate models for both the active, nonlinear circuit and the passive matching and biasing components. Both of these may be problematic to obtain. It is a well-known fact that the amount of *N*th-order distortion is proportional up to *N*th-order derivatives of the I-V and Q-V curves (see [1-3]). Hence, for accurate distortion simulation the I-V and Q-V curves of the active components must be modeled so that not only the dc value, but also the higher order derivatives are correct and continuous (for reference, in early simulation models, already the first derivatives could be discontinuous [4]). Moreover, capacitances are easy to model so that charge is not conserved, which may result in nonphysical rectification and self-biasing in purely capacitive nodes. Hence, especially if the capacitance values depend simultaneously on two terminal voltages, it is important to model the capacitances as charge equations instead of capacitances [5].

The passive components are also tricky to model at RF frequencies, due to their distributed nature at high frequencies. Lossy transmission lines are difficult to model for time-domain simulations, and, in general, the modeling of the passive components tends to be more accurate in the frequency domain. Still, the frequency-domain simulation models of some passive components like step changes in transmission line width may also be inaccurate at higher harmonics. As one extreme of modeling, it is possible to use the measured terminal impedance values of a test circuit in the circuit analysis.

In short, for accurate distortion simulation we require from the simulation models that:

1. The *N*th-order derivatives of the I-V and Q-V curves must be accurate enough for *N*th-order distortion simulations.
2. The frequency responses of the node impedances must be correct up to the highest relevant harmonic. At baseband frequencies, correct modeling of biasing impedances and the thermal impedance are also needed.
3. It would also be very helpful if componentwise information about the dominant distortion sources were available.

4.1.1 Nonlinear Simulation Models

Broadly speaking, there are two types of nonlinear models that PA and transmitter designers use: behavioral black-box models for system simulations, and device models for circuit simulations. These can be further divided based on the modeling technique: models may be analytical, based on some predefined and physics-based, parametrized model functions, or empirical, where the measured data is tabulated and interpolated or modeled with simple splines or polynomials with no clear physical meaning. These main groups are illustrated in Table 4.1.

Behavioral baseband models are widely used for simulating and optimizing entire transmitters and transceivers, and new flavors have been added, for example to model memory effects, as in [6]. However, behavioral models describe either an existing amplifier, or they are used to derive specifications for a so-far-nonexistent amplifier, but they are of limited use in the design of a new power amplifier.

Just for reference, the properties of the most common behavioral models are described here briefly. Plain static AM-AM and AM-PM curves are not capable of modeling memory effects, but models where AM-AM and AM-PM curves depend on the modulation frequency have been

developed. In the Saleh model, linear filters are added both to the input and output side of the AM-AM and AM-PM nonlinear block, as illustrated in Figure 4.1. In the Blum & Jeruchim model, FFT and adequate oversampling is used to find the instantaneous modulation frequency that is used to modify the AM-AM table (both models are described in [7]). Furthermore, Cadence has implemented its own behavioral K-model in its SpectreRF simulator [8]. An example of Volterra type behavioral modeling is called Volterra input output map (VIOMAP). It is conceptually a nonlinear extension of normal *S*-parameters, including harmonic responses, and it has been successfully used in single-tone load-pull simulations [9, 10].

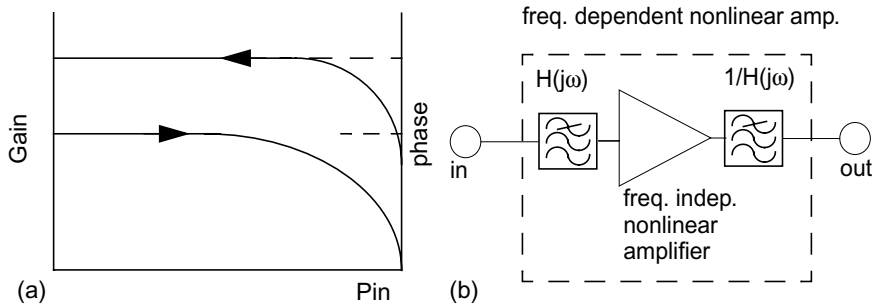


Figure 4.1 (a) AM-AM and AM-PM curves for a power amplifier, and (b) a frequency-dependent nonlinear model based on filters and memoryless nonlinearity. From [11].

Table 4.1
Nonlinear Models for Power Amplifiers

	Behavioral	Device Models
Analytical (physics based)	Saleh Blum-Jeruchim	Mextram VBIC95 MET
Empirical (measurement based)	AM-AM and AM-PM VIOMAP Volterra	Table-based models Volterra

Device models describe the operation of the semiconductor device, and together with proper models for passive and distributed components, models for amplifiers can be built and optimized.

Early semiconductor models were analytical, using equations that were first derived in detail from semiconductor physics and then often simplified to reduce the simulation time. The basic problem with these equation based models is that the chosen functions and control parameters fix the possible shape of I-V and Q-V characteristics, and there may simply not be enough degrees of freedom to model things like I_C-V_{CE} curvature. As an example, in a slightly simplified form the collector current in the basic Gummel-Poon (GP) BJT SPICE model is given by

$$I_C = \frac{2 \cdot IS \cdot \exp\left(\frac{v_{BE}}{V_t}\right) \cdot \left(1 + \frac{v_{CE} - v_{BE}}{VAF}\right)}{1 + \sqrt{1 + \frac{4 \cdot I_S \cdot \exp\left(\frac{v_{BE}}{V_t}\right)}{IKF}}}, \tag{4.1}$$

where the basic exponential I_C-V_{BE} dependency can be modified by only three control parameters: *IS* scales the current, *VAF* (so-called Early voltage [12], illustrated in Figure 4.2) makes an extremely simplified model of the output conductance, and *IKF* (so-called knee current) reduces the gain at high currents [13, 14]. This simple equation covers the entire I-V plane and also fixes the derivatives $d^n I_C/dV^n$, thus fixing the nonlinear behavior. The SPICE GP model can still be used reasonably well for simulating the fundamental signal [15], but especially due to the over-simplified and inherently linear output impedance model, it cannot be used for accurate distortion simulations, as will be seen later in this book.

Better physical models have been developed, and for example Mextram and VBIC for BJTs [16-20] and Motorola MET model [21] for laterally diffused MOS (LDMOS) devices are quite sophisticated. Compared to early SPICE models, they are greatly enhanced. The I-V curve has a more realistic shape and continuous higher order derivatives, charge conserving and continuous capacitance models have been included, and the effects of self-heating have been added into the model. The latter is important for finding the correct dc bias, as the self-heating makes a big difference in the I-V curve. It can also be used to model the thermal memory effects, provided that the thermal model has enough time constants to model both the slowly warming package that affects mainly the dc bias and the microsecond range thermal memory of the chip's surface.

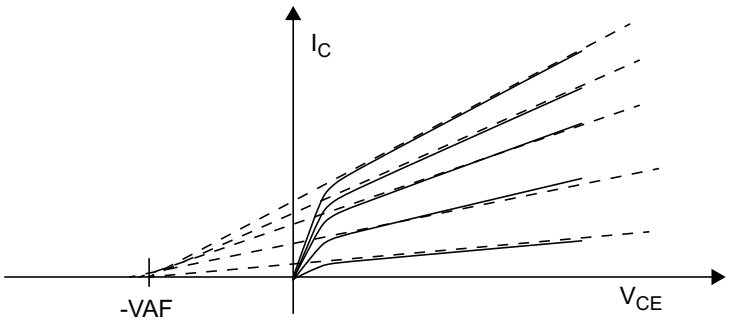


Figure 4.2 Output impedance modeling in a BJT using the Early voltage *VAF*.

Added degrees of freedom in the model increase the model complexity and the number of control parameters. In one extreme, MOS BSIM models have tens of parameters controlling the scaling properties alone. Hence, the complexity of the models tends to grow out of hand and their fitting gets increasingly complicated and sensitive to errors.

Another approach in device modeling is to abandon the equations and use tabulated measured data or fully empirical fitting functions instead. Now any form of I-V and Q-V characteristics can be modeled, and this is the idea behind the Root models, sometimes called "the device knows best" models [5]. There are some technical problems in interpolating tabulated data, as interpolating polynomials easily create oscillations between the data points and hence nonphysical fluctuation in the higher order derivatives. However, tabulated models are very flexible and quite easy to fit as no forcing to the predefined functions is needed.

Volterra models are empirical models as they do not rely on semiconductor physics. The nonlinearities are described as polynomials, the coefficients of which may be obtained either by differentiating physics based I-V and Q-V functions, or by fitting polynomials directly to measured data. The latter approach is used here, and the properties of Volterra models will be studied more deeply in the following section.

4.1.2 The Properties of the Volterra Models

Polynomial models are not automatically quick to simulate; on the contrary, they may converge badly at signal levels higher than the original fitting range. However, the use of polynomial modeling allows the use of a noniterating and efficient Volterra analysis procedure.

However, here the main motivation for using the Volterra simulation technique is not the speed advantage but the fact that it provides an excellent tool for analysis. Dominant distortion mechanisms can be recognized in the same way as in normal ac noise analysis, and owing to the nonlinear analysis, multiple mixing mechanisms can also be recognized, which aids the design of harmonic terminal impedances, for example. Thus, Volterra analysis is one of the few ways of obtaining an understanding of memory effects and aiding design optimization.

Still, polynomial modeling has some shortcomings that need to be recognized. First, polynomial models suffer from the fact that outside the fitting range their response explodes towards infinity. This is the opposite of typical nonlinear modeling functions, where smooth and limited behavior over the entire bias range is a desired property, as it aids convergence and the signal swings are not necessarily a priori known. Hence, the Volterra analysis is not a very general tool. Due to its speed it is used for quick distortion analysis and optimization for low-noise amplifier (LNA) type small-signal circuits in simulators like Voltaire XL [22] and early versions of SPICE, or even as standalone simulators [23]. However, for successful power amplifier analysis, certain preliminary information is needed.

Second, the actual large-signal dc bias voltages are needed in advance. The large signal operation often causes a shift in the dc operating point that affects both the gain and the amount of nonlinearity. This signal-induced dc shift slows the convergence in a harmonic balance simulation, and in the noniterating Volterra calculation procedure it can only be estimated but not completely predicted. To overcome this, we need either to check if the dc shift is significant or to fit the polynomial model at the actual large signal operating point.

Third, during the fitting of the polynomial functions the extent of input and output voltage swings is needed. The real power of the polynomial modeling is that – besides the separation of distortion components – a local fit over just the required voltage range can provide more accurate high-order derivatives than the use of analytical models that have to cover a very broad range of operating regimes with a limited number of control parameters. The larger the fitting range, the less accurate a low-degree polynomial fit may be. Hence, it is desirable to fit along the desired maximum signal excursions, and not much beyond as it compromises the accuracy of polynomial modeling, and also not over a smaller range as the response of the polynomials may then be completely nonphysical outside the fitting range. In that sense, a good guess of the input and output trajectories is needed. Altogether, Volterra analysis is not necessarily an easy-to-use standalone simulation approach, but it provides much

debugging power when used in parallel with other simulation methods such as harmonic balance, for example.

In this book, the examples studied are limited to single-transistor amplifier stages, and the Volterra analysis has been calculated semi-analytically, by deriving symbolically the transfer functions from each distortion source to all node voltages. Symbolic analysis is by no means necessary and limits the analysis to fixed CE or CS amplifier architectures and two-tone test signal. Instead, the Volterra analysis can – with almost the same resolution of distortion components achieved here – be calculated fully numerically by recursively using ordinary linear ac analysis on any circuit described by a standard modified nodal analysis (MNA) matrix representation and nonlinear current sources. To aid the study of more complicated amplifier topologies, the effects of multidimensionally controlled charges, and other complicated features, a fifth-order, multitone numerical simulator with a simple netlist interface has been designed [24].

4.2 Nonlinear I-V and Q-V Characteristics

Most transistor models are based on either a Π or T model. Here, a Π model is used, and this section presents the typical conductive (I-V) and capacitive (Q-V type) nonlinearities appearing in the Π model of BJT, heterojunction BJT (HBT) and field-effect transistor (FET) devices. The BJT is used as an example, but the same models with different sets of polynomial coefficients can be used for FET transistors as well.

As explained before, the Volterra model is based on polynomial modeling of I-V and Q-V curves. Measuring these may be somewhat complicated, as seen in more detail in Chapter 5. Charge as such cannot be measured directly, and we must rely on extracting the capacitances from ac measurements and integrating the charge equations from the obtained capacitance values. In a similar manner, the I-V curve can be mostly reconstructed from g_m and g_o values obtained from *S*-parameter measurements, but the actual I-V curve is a safer starting point.

The models presented are electrothermal models, which means that the junction temperature also appears as a free variable. However, the dc temperature rise is included in the bias point, and only the temperature variations caused by dynamic self-heating are considered. As the dissipated power is a product of voltage and current, the fundamental assumption in the following is that the ac component in the junction temperature is already a second-order phenomenon. Hence, a third-order model will contain only the first power of the temperature, which also means that the temperature dependencies of circuit elements are considered linear.

4.2.1 I_C-V_{BE}-V_{CE} Characteristic

In most reported BJT/HBT Volterra series analyses the collector current is considered a function of base voltage only [25-27], which captures the dominant exponential input-output nonlinearity but assumes that the output conductance is constant. In MESFET Volterra series analysis the effects of drain voltage are usually implemented by a polynomial for g_o (see [28]), but even that does not capture all nonlinearities.

Already the simple expression in (4.1) is a three-dimensional function of V_{BE}, V_{CE}, and junction temperature T, as $V_t = kT/q$. A polynomial model is derived simply by expanding the large-signal I-V function

$$I_C = f(V_{BE}, V_{CE}, T), \tag{4.2}$$

to a three-input Taylor series around the dc operating point V_{BEQ}, V_{CEQ}, T_Q. Hence, an electrothermal third-degree series expansion of the ac current can be written as

$$\begin{aligned} i_c = {} & g_m v_{be} + K_{2GM} v_{be}^2 + K_{3GM} v_{be}^3 \\ & + g_o v_{ce} + K_{2GO} v_{ce}^2 + K_{3GO} v_{ce}^3 \\ & + K_{2GMGO} v_{be} v_{ce} + K_{3GM2GO} v_{be}^2 v_{ce} + K_{3GMGO2} v_{be} v_{ce}^2 \\ & + K_{2GMT} t_J + K_{3GMT} t_J v_{be} + K_{3GOT} t_J v_{ce} \end{aligned} \tag{4.3}$$

where $v_{be} = v_{BE}$-V_{BEQ}, v_{ce}=v_{CE}-V_{CEQ} and $t_J = T$-T_Q, and K_{ixxx} is the ith degree nonlinearity coefficient of element xxx (alternatively, K_{MNP} could be used to mark a term $v_{be}^M v_{ce}^N t_J^P$). The first row models the effect of v_{be} alone, and the second row the effects of v_{ce} alone (i.e., a nonlinear output conductance). These are not sufficient, though, as cross-products of v_{be} and v_{ce} may also appear, and these are listed in row 3. Finally, the temperature change t_J contributes to the current, and it, too, may mix with both terminal voltages, causing the last three terms on row 4.

The effects of the electrical nonlinearities are demonstrated in Figure 4.3, where the collector current at three base voltages is plotted as a function of collector voltage at three different base voltages. If all the coefficients except the g_m are zero, we obtain the three equally spaced horizontal lines shown in Figure 4.3(a). Since the lines are exactly horizontal, the output conductance is zero and the collector voltage does not affect the amount of current. Furthermore, since the lines are equally spaced, the transconductance is linear. However, if K_{2GM} or K_{3GM} deviates

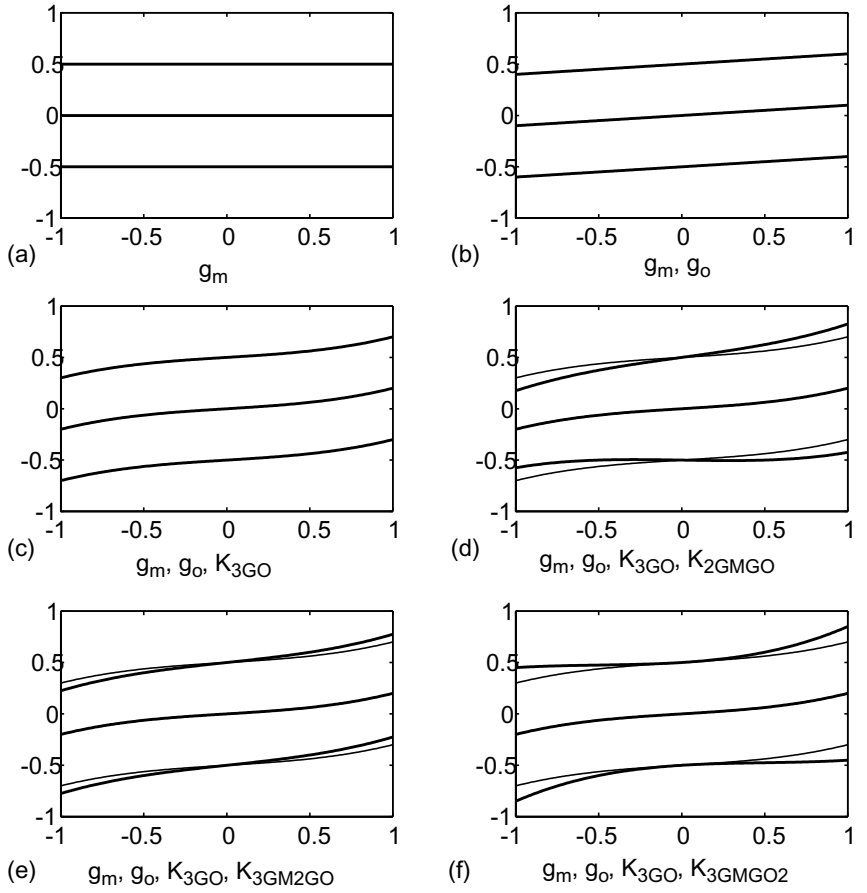


Figure 4.3 Demonstration of collector current nonlinearities. Vertical axis is the collector current and horizontal axis V_{CE} voltage. (a) Linear response, (b) non-zero g_o, (c) non-zero K_{3GO}, (d) non-zero K_{2GMGO}, (e) non-zero K_{3GM2GO}, and (f) non-zero K_{3GMGO2}.

from zero, the lines in a I-V plane become unequally spaced, indicating a nonlinear transconductance.

The effects of g_o are demonstrated in Figure 4.3(b), in which just the linear terms g_m and g_o exist. Compared to Figure 4.3(a) the lines now have a nonzero slope that is proportional to g_o and independent of v_{be}.

While Figure 4.3(b) is still fully linear, the nonlinearity of the output conductance is demonstrated in Figure 4.3(c), where the slope of the current varies with V_{CE}. In this case only K_{3GO} has a nonzero value in

Figure 4.3(c), but both K_{2GO} and K_{3GO} can be used to model curvature effects of the output conductance such as saturation and breakdown.

Figure 4.3(d-f) illustrates the effects of the cross-terms K_{2GMGO}, K_{3GM2GO} and K_{3GMGO2}, respectively, that model the interaction between base and collector nonlinearities. To aid comparison, the thin lines in Figure 4.3(d-f) are copied from (c) where only g_m, g_o and K_{3GO} have nonzero values. K_{2GMGO} (corresponding to the $v_{be}v_{ce}$ term) is nonzero in Figure 4.3(d), and as a result of it the slope of the lines varies not only as a function of collector voltage as in Figure 4.3(c), but also with the base voltage. This is essentially needed to model the Early effect illustrated in Figure 4.2. Similar reasoning applies to K_{3GMGO2} and K_{3GM2GO}, that shape the output conductance as functions of $v_{be}v_{ce}$ and v_{be}^2, respectively, as shown in Figure 4.3(e, f).

The I-V curves of different modeling approaches are further compared in Figure 4.4. If the collector current is modeled as a one-dimensional function of base voltage and linear g_o, just straight line I-V curves are generated, as seen from the thin dashed line in Figure 4.4. The I-V curves simulated using the SPICE Gummel-Poon model are also straight (and thin) lines, but their slope and hence the output conductance vary with the collector current, as suggested in Figure 4.2. In reality, however, BJT I-V curves are far from straight lines under large or semi-large signal conditions, due to quasi-saturation and breakdown effects. The curvature can be modeled by using one-dimensional polynomials for both v_{be} and v_{ce}, as illustrated by the thick solid line in Figure 4.4. However, elementary phenomena such as the Early effect cannot be modeled without the cross-terms that cause the I_C-V_{CE} slope to depend on the value of V_{BE}. This is illustrated by the thick dashed line, corresponding to a full series expansion (4.3). The shapes of the saturation and breakdown also depend on the base voltage, and this makes the use of cross-terms mandatory to avoid large errors at the corners of the I-V plane, as seen in Figure 4.4.

Finally, the electrothermal effects of the collector current are discussed, modeled by the last three terms in (4.3). A second-degree term K_{2GMT} models a temperature-dependent shift in the current, as illustrated in Figure 4.5(a, b). It is worth noting that K_{2GMT} cannot be derived from the small-signal parameters g_m and g_o; instead, actual current measurements are needed. K_{3GOT} is a third-degree term that includes a combined effect of temperature and collector voltage as indicated in Figure 4.5(c). Essentially, it models the temperature dependency of the output conductance. Similarly, K_{3GMT} models the combined effects of temperature and base voltage, visualized in Figure 4.5(d). Since the slope of that curve describes the transconductance, K_{3GMT} can be considered a change in the transconductance as a function of temperature.

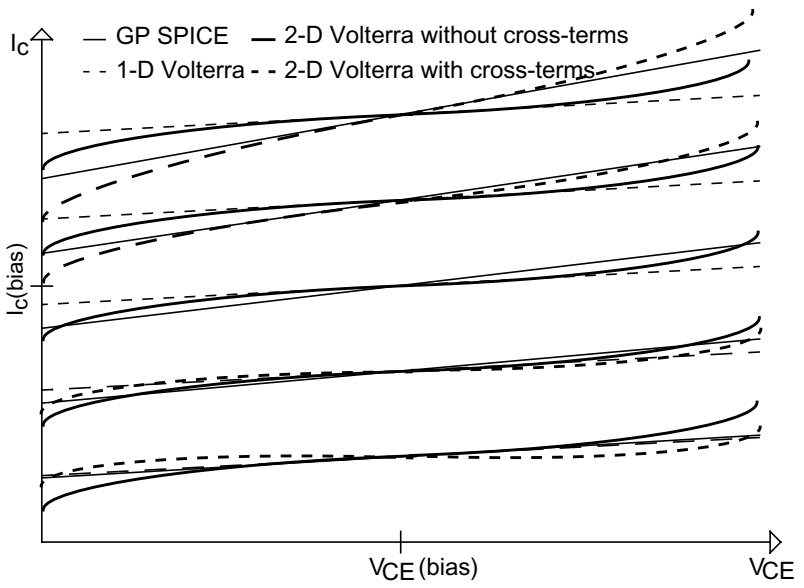


Figure 4.4 I-V characteristics of three Volterra models and Gummel-Poon SPICE model. From [11].

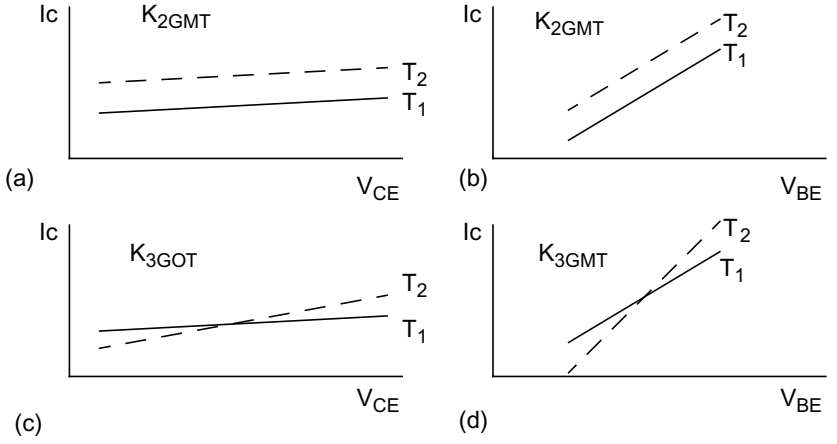


Figure 4.5 The effects of electrothermal nonlinearity coefficients. Non-zero K_{2GMT} on (a) V_{CE}-I_C and (b) V_{BE}-I_C axis. Effects of non-zero (c) K_{3GOT} and (d) K_{3GMT}.

4.2.2 g_{pi} and r_{bb}

The I_D-V_{DS} characteristic is the only important conductive nonlinearity in FET-type transistors. In BJTs, two other conductive nonlinearities exist: The I_B-V_{BE} nonlinearity caused by exponential g_{pi} and the nonlinear r_{bb}. The effect of g_{pi} conductance is usually more important, and it is also easier to model. In theory the I_B-V_{BE} equation should be roughly (4.1) divided by the current gain β, but some simplifications can be made. Since the base current does not depend strongly on the collector voltage, we can use a two-dimensional model of v_{be} and junction temperature t_J dependence only:

$$i_b = g_{pi} \cdot v_{be} + K_{2GPI} \cdot v_{be}^2 + K_{3GPI} \cdot v_{be}^3 + K_{2GPIT} \cdot t_J + K_{3GPIT} \cdot t_J \cdot v_{be} \tag{4.4}$$

Here, the coefficients have similar meanings as before. The linear term is modeled by g_{pi}, and K_{2GPI} and K_{3GPI} model its exponential curvature. Again, K_{2GPIT} models the I_B shift caused by self-heating, and K_{3GPIT} can be seen as a temperature dependence of the linear g_{pi} term.

The intrinsic base resistance r_{bb} is a bit trickier to model. It is a series resistance between the intrinsic and extrinsic base points, but its value depends on the current crowding in the base area and also on the value of the intrinsic v_{be}. Thus, it must be modeled as a three-dimensional conductance, being controlled by the voltage across the resistor (v_{bb}=v_{bext}-v_{bint}), the intrinsic base voltage v_{beint}, and the junction temperature t_J. All the v_{beint}^k terms (k=1,2,...) are zero, but the current crowding effect is modeled with the cross-terms between v_{bb} and v_{beint}, as shown in (4.5). However, r_{bb} is usually small and it has been modeled simply as a linear conductance in the following examples.

$$i_{rbb} = g_{bb} \cdot v_{bb} + K_{211} \cdot v_{bb} \cdot v_{beint} + \ldots \tag{4.5}$$

4.2.3 Capacitance Models

As explained before, the capacitances are modeled as polynomial charges that are then differentiated with respect to time to get the displacement current. The charge may be – and often is – controlled by more than one port voltage, in which case a multidimensional polynomial similar to (4.3) must be used. The charge may also be modeled as a transcapacitance, and in this case the charge does not appear between the controlling nodes but in

some other node. In the following examples, only one controlling voltage is assumed, and as an example, (4.6) represents the base-to-emitter charge as functions of base-to-emitter voltage and temperature.

$$Q_{be} = C_{\mathrm{pi}} \cdot v_{\mathrm{be}} + K_{2\mathrm{CPI}} \cdot v_{\mathrm{be}}^2 + K_{3\mathrm{CPI}} \cdot v_{\mathrm{be}}^3 + K_{2\mathrm{CPIT}} \cdot t_{\mathrm{J}} + K_{3\mathrm{CPIT}} \cdot t_{\mathrm{J}} \cdot v_{\mathrm{be}} \tag{4.6}$$

From this equation, the corresponding measurable capacitance C_{pi} and the nonlinear current source i_{NLCPI} are obtained simply by differentiating the charge equation (4.6) with respect to v_{be} and time, respectively.

$$c_{pi}(v_{be}) = C_{\mathrm{pi}} + 2K_{2\mathrm{CPI}} \cdot v_{\mathrm{be}} + 3K_{3\mathrm{CPI}} \cdot v_{\mathrm{be}}^2 + K_{3\mathrm{CPIT}} \cdot t_{\mathrm{J}}. \tag{4.7}$$

$$i_{\mathrm{NLCPI}} = j\omega \cdot (C_{\mathrm{pi}} \cdot v_{\mathrm{be}} + K_{2\mathrm{CPI}} \cdot v_{\mathrm{be}}^2 + K_{3\mathrm{CPI}} \cdot v_{\mathrm{be}}^3 + K_{2\mathrm{CPIT}} \cdot t_{\mathrm{J}} + K_{3\mathrm{CPIT}} \cdot t_{\mathrm{J}} \cdot v_{\mathrm{be}}) \tag{4.8}$$

In (4.8), ω is simply the frequency of the distortion tone; hence, capacitances do not cause dc distortion currents but distort most heavily at the harmonic frequencies. Equation (4.7) shows that the temperature-dependent charge term $K_{2\mathrm{CPIT}}$ cannot be derived from capacitance measurements; still, a time-varying junction temperature may cause a current proportional to it. Otherwise, the first term C_{pi} in (4.6) describes the small-signal capacitance, and $K_{2\mathrm{CPI}}$ and $K_{3\mathrm{CPI}}$ define its electrical nonlinearity. $K_{3\mathrm{CPIT}}$ describes the charge being a function of both the controlling voltage and junction temperature and since $C = \mathrm{d}Q/\mathrm{d}v$, the effects of $K_{3\mathrm{CPIT}}$ can be seen as the temperature dependency of the capacitance value.

As seen from (4.6), a *linear* C-V trend $K_{2\mathrm{CPI}}$ causes quadratic charge nonlinearity. Similarly, a capacitance proportional to v^2 ($K_{3\mathrm{CPI}}$) causes cubic nonlinearity. Different types of capacitances have different characteristics, as illustrated in Figure 4.6(a). The base-emitter capacitance C_{pi} is exponential [see (4.27)] and therefore highly nonlinear. Reverse-biased p-n or Schottky junctions seen in BJTs and in FETs are just weakly nonlinear, and they can be linearized further by increasing the reverse bias. MOSFET-type transistors have peculiar gate capacitances and, for example, the C_{GS} makes a clear dip around the threshold voltage. If the MOSFET is operated close to cutoff, this dip can cause a large amount of second-degree nonlinearity.

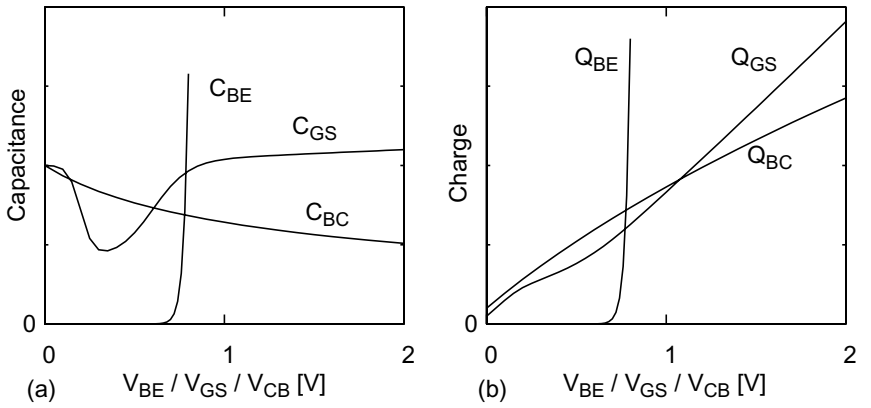


Figure 4.6 Normalized (a) capacitances and (b) charges of BJT C_{BE}, MOSFET C_{GS}, and BJT C_{BC}.

4.3 Model of a Common-Emitter BJT/HBT Amplifier

The direct method is now used to calculate the IM3 components in a common-emitter BJT/HBT amplifier, using the procedure outlined in Section 2.5.2. The analysis starts by building a model for the circuit, from which the fundamental amplitudes are found by a linear ac analysis. Then, the second-order currents and voltages and third-order currents and voltages are calculated using the procedure shown in Section 4.3.2.

4.3.1 Linear Analysis

The model of a common-emitter BJT amplifier, shown in Figure 4.7, includes the input impedance Z_{IN} (lumping both the matching network and bias circuitry), base-emitter conductance g_{pi} and capacitance C_{pi}, feedback capacitance C_{BC}, output capacitance C_{CE} and output conductance g_o, transconductance g_m, load impedance Z_L, and emitter impedance Z_E. The input and load impedances include not only the impedances of the matching networks, but also the impedances of the bias networks and package parasitics, and Z_{IN} further consists of the output impedance of the preceding stage and the intrinsic base resistance r_{bb}, as shown in Figure 4.7.

To reduce the amount of equations, the input voltage source can be replaced by its Norton equivalent source

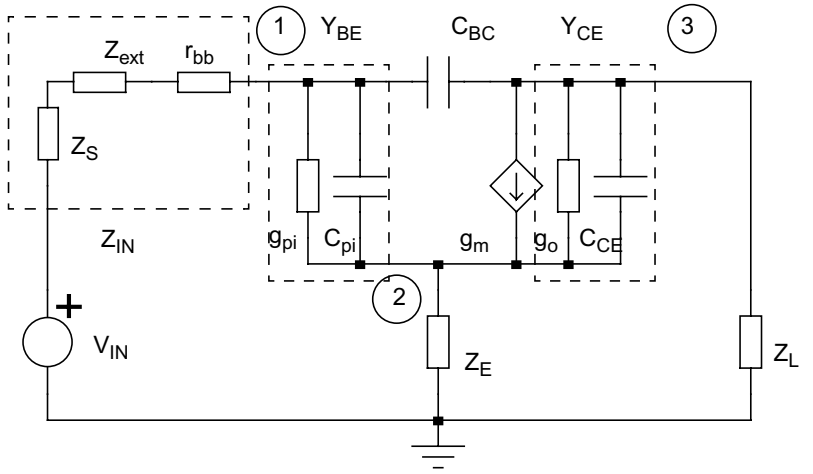


Figure 4.7 Linearized first-order circuit of a common-emitter BJT amplifier.

$$i_{\mathrm{IN}} = Y_{\mathrm{IN}}(s) \cdot v_{\mathrm{IN}}, \tag{4.9}$$

and using the following shorthand notations

$$\begin{aligned} Y_{\mathrm{BE}}(s) &= g_{\mathrm{pi}} + sC_{\mathrm{pi}} \\ Y_{\mathrm{CE}}(s) &= g_o + sC_{\mathrm{CE}}. \\ Y_{\mathrm{BC}}(s) &= sC_{\mathrm{BC}} \end{aligned} \tag{4.10}$$

The matrix equation can be set up by changing all the impedances in Figure 4.7 to admittances and applying Kirchhoff's current law at nodes 1 to 3. This results in

(4.11)

$$\begin{bmatrix} i_{\mathrm{IN}} \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} Y_{\mathrm{IN}} + Y_{\mathrm{BE}} + Y_{\mathrm{BC}} & -Y_{\mathrm{BE}} & -Y_{\mathrm{BC}} \\ -g_{\mathrm{m}} - Y_{\mathrm{BE}} & g_{\mathrm{m}} + Y_{\mathrm{E}} + Y_{\mathrm{BE}} + Y_{\mathrm{CE}} & -Y_{\mathrm{CE}} \\ g_{\mathrm{m}} - Y_{\mathrm{BC}} & -g_{\mathrm{m}} - Y_{\mathrm{CE}} & Y_{\mathrm{L}} + Y_{\mathrm{CE}} + Y_{\mathrm{BC}} \end{bmatrix} \cdot \begin{bmatrix} v_B \\ v_E \\ v_C \end{bmatrix}$$

and the responses to i_{IN} can be found using Cramer's rule. Thus, the voltages at the base, emitter, and collector are given by

$$v_B(s) = \begin{pmatrix} Y_{BE}Y_{CE} + g_m Y_L + Y_{BE}Y_L + Y_{CE}Y_L + Y_{CE}Y_E \\ + Y_L Y_E + Y_{BC}(Y_{BE} + Y_{CE} + Y_E + g_m) \end{pmatrix} \cdot \frac{i_{IN}(s)}{det(s)} \quad (4.12)$$

$$v_E(s) = \frac{Y_{BE}Y_{CE} + g_m Y_L + Y_{BE}Y_L + Y_{BC}(Y_{BE} + Y_{CE} + g_m)}{det(s)} \cdot i_{IN}(s) \quad (4.13)$$

and

$$v_C(s) = \frac{Y_{BE}Y_{CE} - g_m Y_E + Y_{BC}(Y_{BE} + Y_{CE} + Y_E + g_m)}{det(s)} \cdot i_{IN}(s) \quad (4.14)$$

where the determinant of the admittance matrix is written as

$$\begin{aligned} det(s) &= Y_{BE}Y_{CE}(Y_L + Y_E + Y_{IN}) + Y_{IN}Y_L \\ &\cdot (Y_{BE} + Y_{CE} + g_m + Y_E) + Y_{CE}Y_E Y_{IN} + Y_{BE}Y_L Y_E \\ &+ Y_{BC} \cdot [Y_{BE}Y_{IN} + Y_{CE}Y_{IN} + Y_E Y_{IN} + Y_{BE}Y_E \\ &+ Y_{BE}Y_L + Y_{CE}Y_L + Y_E Y_L + Y_{CE}Y_E + g_m Y_L \\ &+ g_m Y_{IN} + g_m Y_E] \end{aligned} \quad (4.15)$$

From these, the base-emitter and collector-emitter voltages are simply $v_{BE}(s) = v_B(s) - v_E(s)$ and $v_{CE}(s) = v_C(s) - v_E(s)$, respectively. Finally, as both v_{BE} and v_{CE} are frequently used in calculating the distortion generated by the g_m element, for example, it is handy to derive their ratio

$$TF(s) = \frac{v_{CE}(s)}{v_{BE}(s)}. \quad (4.16)$$

The purpose of the linear analysis is to obtain the fundamental voltage amplitudes across all nonlinear components so that we can proceed in calculating the nonlinear currents generated in these components. Before doing that, a few observations about the signal swings are worth making.

The exponential response of a BJT is extremely nonlinear and it does not tolerate higher than 10 to 30 mV signal amplitude in the BE junction without distorting excessively. That does not sound like very much for a

power amplifier, but a couple of things happen to help the situation. First, the device is not completely exponential, but when driven to high injection, a BJT linearizes considerably as modeled by the parameter *IKF* in (4.1). Second, the amplifier has several feedback mechanisms that reduce the signal level in the BE junction. The series emitter impedance causes a linearizing series feedback, and C_{BC} causes a shunt feedback. The effect of C_{BC} is very important, as the strong capacitive feedback considerably lowers the impedance at the base, and so reduces the BE voltage swing and the amount of generated distortion for a given driving power.

4.3.2 Nonlinear Analysis

In this section the nonlinear model of a CE BJT amplifier is presented and the equations for IM3 distortion are derived. The circuit has three two-input and one three-input I-V and Q-V nonlinearities, modeled by 27 first-, second-, and third-degree coefficients, of which 18 are purely electrical (C_{pi}, K_{2CPI}, K_{3CPI}, C_{bc}, K_{2CBC}, K_{3CBC}, g_{pi}, K_{2GPI}, K_{3GPI}, g_m, K_{2GM}, K_{3GM}, g_o, K_{2GO}, K_{3GO}, K_{2GMGO}, K_{3GM2GO}, and K_{3GMGO2}) and nine are related to dynamic temperature variations (K_{2CPIT}, K_{3CPIT}, K_{2CBCT}, K_{3CBCT}, K_{2GPIT}, K_{2GPIT}, K_{2GMT}, K_{3GMT}, and K_{3GOT}).

Eventually, this analysis will present the IM3 tone as a vector sum of seven terms caused by cubic electrical nonlinearities, 21 terms caused by cascaded quadratic nonlinearities where the rectified envelope information is upconverted to IM3, 21 similar terms where the second harmonics are mixed down to IM3, and finally, five cubic and 24 cascaded second-degree electrothermal terms. This resolution may seem excessive, but it was chosen to illustrate the real multitude of different mechanisms that generate distortion; it also clearly illustrates that very much information is lost if only the effects of the cubic terms are analyzed. With larger circuits and higher order analysis it will be necessary to compress the data somehow, but the principle is still the same: we want to see how much of the total IM3 is caused by mixing distortion from the dc or harmonic bands and to be able to minimize the total distortion (or its memory effects), we want to see what nonlinearities and impedances actually are causing the distortion voltages at these harmonic bands.

Due to the multitude of terms, not all of them will be discussed separately. The full analysis is shown in Appendix C, and next, the calculation procedure is illustrated with some examples.

4.3.2.1 Second-Order Distortion Currents

The circuit for solving the second-order responses is shown in Figure 4.8, where the linear input voltage shown in Figure 4.7 is now short-circuited and the second-order distortion current sources i_{NL2XX} are added in parallel to all nonlinear circuit elements. Currents ending with T are electrothermal currents that are discussed in Section 4.3.2.5. As before, Z_{IN}, Z_E, and Z_L lump the package parasitics, and biasing and matching impedances.

For calculating the self-heating effects, the instantaneous power dissipation is calculated as $P_{diss} = v_{CE}i_C$, and the thermal impedance Z_{TH} shown in Section 3.4 is used to calculate the instantaneous temperature fluctuation $T_J(\omega_2–\omega_1)$ at frequency $\omega_2–\omega_1$. It is possible to use different temperatures for different circuit elements, but as they all are physically located close to the base area, a common temperature is used here. For large devices, however, it may be advantageous to split the device into smaller parallel devices that may see different temperature variations.

The I_C-V_{BE}-V_{CE} nonlinearity is modeled as a three-dimensional function of the v_{BE}, v_{CE}, and temperature, including g_m and g_o nonlinearity

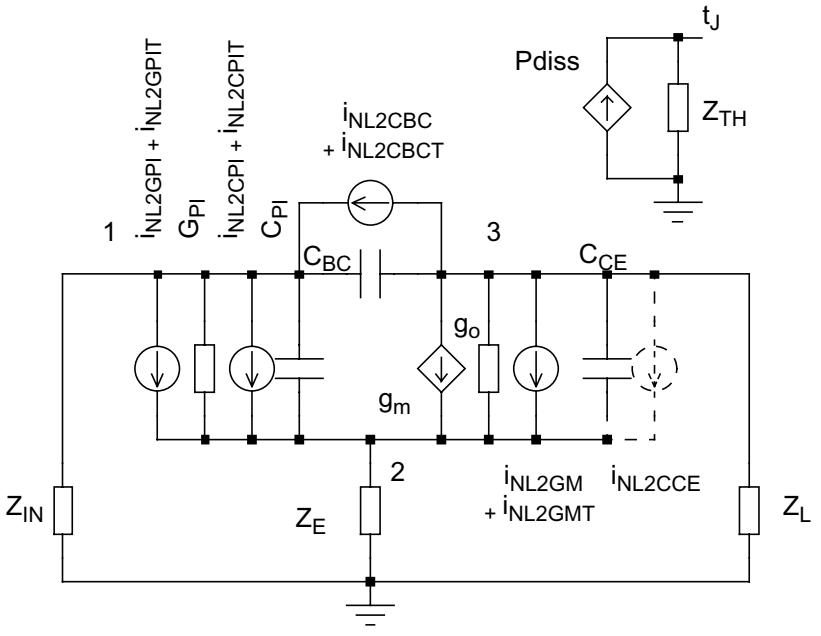


Figure 4.8 Representation of a circuit containing current sources for second-order responses.

and all up to third-degree cross-terms. The g_{pi} and C_{pi} are nonlinear functions of the base-emitter voltage and temperature, and the weakly nonlinear C_{BC} is controlled by the collector-base voltage and temperature.

We start the analysis by calculating the second-order distortion currents i_{NL2XX}. As an example, the second-order envelope current at ω_2–ω_1 caused by the g_{pi} is, using Table 2.5,

$$i_{NL2GPI}(\omega_2 - \omega_1) = K_{2GPI} V_{BE}(\omega_2) \overline{V_{BE}(\omega_1)}. \tag{4.17}$$

As another example, the second-order envelope current caused by the I_C-V_{BE}-V_{CE} nonlinearity (g_m) is

$$\begin{aligned} i_{NL2GM}(\omega_2 - \omega_1) &= K_{2GM} V_{BE}(\omega_2) \overline{V_{BE}(\omega_1)} \\ &+ K_{2GO} V_{CE}(\omega_2) \overline{V_{CE}(\omega_1)} \\ &+ 0.5 \cdot K_{2GMGO}(V_{BE}(\omega_2) \overline{V_{CE}(\omega_1)} + V_{CE}(\omega_2) \overline{V_{BE}(\omega_1)}) \end{aligned} \quad , \tag{4.18}$$

which combines the effects of the second-degree input nonlinearity $K_{2GM}*v_{BE}^2$, output nonlinearity $K_{2GO}*v_{CE}^2$ and the input-output cross-term $K_{2GMGO}*v_{BE}*v_{CE}$, all seen in the I-V model (4.3).

As noted from Table 2.5, the selection of the frequencies of the phasors and the values of possible constant terms depend on the tone frequency: a product $V_{BE}(\omega 1)*V_{BE}(\omega 1)$ results in a tone at $2\omega 1$, for example. Above, the tones of the phasors are chosen so that they always result in distortion at the envelope frequency ω_2–ω_1. The phasors V_{BE} and V_{CE} for the fundamental tones ω_1 and ω_2 are calculated using (4.12)-(4.14).

4.3.2.2 Transimpedance Transfer Functions and Second-Order Voltages

Next, we need to convert the distortion currents to distortion voltages in various nodes. Here, a semisymbolic analysis has been chosen, so that the transfer functions TF_{XYZ} from a nonlinear current source between nodes X and Y to a node voltage Z have been derived by hand. The general idea is that within each nonlinear element, the amplitudes of the distortion currents are calculated using the lower order voltage phasors, and using the transfer functions, the generated distortion currents are then converted to distortion voltages in the chosen nodes. The transfer functions can be derived from (4.11) by replacing i_{IN} with a test current source between nodes X and Y. As an example, the transfer function TF_{BEB} is of the form

$$TF_{\mathrm{BEB}}(s) = \frac{V_{\mathrm{B}}}{i_{BE}} = \frac{-[Y_{\mathrm{E}} \cdot Y_{\mathrm{L}} + Y_{\mathrm{CE}} \cdot Y_{\mathrm{E}} + Y_{\mathrm{CE}} \cdot Y_{\mathrm{L}} + g_{\mathrm{m}} \cdot Y_{\mathrm{L}} + Y_{\mathrm{BC}} \cdot Y_{\mathrm{E}}]}{det(s)} \quad (4.19)$$

where *det(s)* is given by (4.15). A complete set of transfer functions is shown in Appendix C. Using these notations, the complete second-order envelope voltage at the base node is now given by

$$\begin{aligned} &V_{\mathrm{B2}}(\omega_2 - \omega_1) \\ &= TF_{\mathrm{BEB}}(\omega_2 - \omega_1) \cdot (i_{\mathrm{NL2GPI}} + i_{\mathrm{NL2CPI}} + i_{\mathrm{NL2GPIT}} + i_{\mathrm{NL2CPIT}}) \\ &+ TF_{\mathrm{CEB}}(\omega_2 - \omega_1) \cdot (i_{\mathrm{NL2GM}} + i_{\mathrm{NL2GMT}}) \\ &+ TF_{\mathrm{CBB}}(\omega_2 - \omega_1) \cdot (i_{\mathrm{NL2CBC}} + i_{\mathrm{NL2CBCT}}) \end{aligned} \quad (4.20)$$

where each distortion current source is multiplied by the proper transimpedance that converts the current to a voltage at the base. Note also that there are eight different current sources (four of them being electrothermal and marked with *T*) and each one of them will have an equation resembling (4.17) or (4.18).

To solve the second harmonic voltages, we need to rewrite the current equations of the i_{NLXX} sources for the desired harmonic, and to recalculate (4.20) at that frequency. After this procedure, we have the base envelope and second harmonic tones given as sums of eight different contributions. By storing them separately, we can see which one is dominant, and which ones may cancel each other.

4.3.2.3 Solving the Third-Order Voltages

Next, we are ready to calculate the IM3 collector voltage at frequency $2\omega_1-\omega_2$. The third-order analysis uses exactly the same equivalent circuit as the second-order analysis, and the collector voltage is given by (4.21)

$$\begin{aligned} &V_{\mathrm{C3}}(2\omega_1 - \omega_2) \\ &= TF_{\mathrm{BEC}}(2\omega_1 - \omega_2) \cdot (i_{\mathrm{NL3GPI}} + i_{\mathrm{NL3CPI}} + i_{\mathrm{NL3GPIT}} + i_{\mathrm{NL3CPIT}}) \\ &+ TF_{\mathrm{CEC}}(2\omega_1 - \omega_2) \cdot (i_{\mathrm{NL3GM}} + i_{\mathrm{NL3GMT}}) \\ &+ TF_{\mathrm{CBC}}(2\omega_1 - \omega_2) \cdot (i_{\mathrm{NL3CBC}} + i_{\mathrm{NL3CBCT}}) \end{aligned} \quad (4.21)$$

where the distortion currents are now calculated for tone $2\omega_1$–ω_2, and the transfer functions TF_{XYC} (also calculated at that frequency) are used to convert the nonlinear currents i_{NL} from port X-Y to collector voltage. However, the equations of the nonlinear current sources will be messier than in the second-order analysis, because IM3 currents are not only caused by cubic nonlinearities, but by cascaded second-degree nonlinearities, as well. According to Table 2.6, both the envelope voltage and the second harmonic appear in the IM3 current caused by the nonlinear I_{B}-V_{BE}, for example,

$$i_{\mathrm{NL3GPI}}(2\omega_1 - \omega_2) = \frac{3}{4}K_{3\mathrm{GPI}}V_{BE}(\omega_1)^2\overline{V_{BE}(\omega_2)} + K_{2\mathrm{GPI}}V_{BE}(\omega_1)\overline{V_{BE}(\omega_2 - \omega_1)} + K_{2\mathrm{GPI}}V_{BE}(2\omega_1)\overline{V_{BE}(\omega_2)} \tag{4.22}$$

Similar responses are caused for example by the coefficient pairs $K_{2\mathrm{GM}}$, $K_{3\mathrm{GM}}$, and $K_{2\mathrm{GO}}$, $K_{3\mathrm{GO}}$ of the I_{C}-V_{BE}-V_{CE} nonlinearity. In addition to these, the input-output cross-terms $K_{2\mathrm{GMGO}}$, $K_{3\mathrm{GM2GO}}$, and $K_{3\mathrm{GMGO2}}$ cause the following additional terms to $i_{\mathrm{NL3GM}}(2\omega_1–\omega_2)$

$$\begin{aligned} &1/2 \cdot K_{2\mathrm{GMGO}} \cdot [\overline{V_{\mathrm{BE}}(\omega_2)}V_{\mathrm{CE}}(2\omega_1) + V_{\mathrm{BE}}(\omega_1)\overline{V_{\mathrm{CE}}(\omega_2 - \omega_1)} \\ &+ \overline{V_{\mathrm{BE}}(\omega_2 - \omega_1)}V_{\mathrm{CE}}(\omega_1) + V_{\mathrm{BE}}(2\omega_1)\overline{V_{\mathrm{CE}}(\omega_2)}] \\ &+ 1/4 \cdot K_{3\mathrm{GM2GO}} \\ &\cdot [2\overline{V_{\mathrm{BE}}(\omega_2)}V_{\mathrm{BE}}(\omega_1)V_{\mathrm{CE}}(\omega_1) + V_{\mathrm{BE}}(\omega_1)^2\overline{V_{\mathrm{CE}}(\omega_2)}] \\ &+ 1/4 \cdot K_{3\mathrm{GMGO2}} \\ &\cdot [2V_{\mathrm{BE}}(\omega_1)\overline{V_{\mathrm{CE}}(\omega_2)}V_{CE}(\omega_1) + \overline{V_{\mathrm{BE}}(\omega_2)}V_{\mathrm{CE}}(\omega_1)^2] \end{aligned} \tag{4.23}$$

where all such combinations of linear and second-order input and output voltages are shown that their products make a tone at $2\omega_1$-ω_2. Note that all possible permutations are needed in the cross-terms. For example, the last term in (4.23) consists of one V_{BE} and two V_{CE} voltages at frequencies ω_1, ω_1, and $–\omega_2$, and they can be arranged in the following three combinations: ($V_{\mathrm{BE}}(\omega_1)$, $V_{\mathrm{CE}}(\omega_1)$, $V_{\mathrm{CE}}(–\omega_2)$), ($V_{\mathrm{BE}}(\omega_1)$, $V_{\mathrm{CE}}(–\omega_2)$, $V_{\mathrm{CE}}(\omega_1)$), and ($V_{\mathrm{BE}}(–\omega_2)$, $V_{\mathrm{CE}}(\omega_1)$, $V_{\mathrm{CE}}(\omega_1)$), the first two giving the same result. Numerical spectral convolution gives all these combinations automatically.

Now, IM3 is described in terms of fundamental and second-order node voltages. Next, we take one step further and write the distortion as a function of fundamental input voltages only. This complicates the expressions quite a lot but has the major benefit that it maintains the information of the origin of second-order distortion voltages.

4.3.2.4 IM3 Shown as a Function of Fundamental Input Tones Only

To present the distortion as a function of input tones only, we need to do two things. First, $V_{CE}(j\omega)$ is written as $TF(j\omega)V_{BE}(j\omega)$, where TF is given in (4.16) and in Appendix C. Second, the second-order tones must also be written in terms of the fundamental inputs, which results in altogether 42 cascaded second-degree terms, all listed in Appendix C.

The purely cubic portion of (4.21) is written in (4.24) in terms of V_{BE} only. Here, the nonlinear currents are written according to Table 2.6 as functions of $V_{BE}(j\omega)$ and $TF(j\omega)V_{BE}(j\omega)$. For example, g_o in the first term in (4.24) is controlled by the product $V_{CE}(\omega 1)^2 V_{CE}(-\omega 2)$. Input nonlinearities g_m, g_{pi}, and C_{pi} are controlled by V_{BE} values only, and the cross-terms are controlled both with input and output voltages. For C_{BC}, the voltage V_{CB} must be expressed as $V_{CE}-V_{BE} = (TF-1)V_{BE}$.

$$
\begin{aligned}
&V_{C3CUBIC}(2\omega_1-\omega_2) \\
&= \{TF_{CEC}(2\omega_1-\omega_2)\cdot(3/4\cdot K_{3GO}\cdot TF(\omega_1)^2\cdot\overline{TF(\omega_2)}) \\
&+ TF_{CEC}(2\omega_1-\omega_2)\cdot(3/4\cdot K_{3GM}) \\
&+ TF_{BEC}(2\omega_1-\omega_2)\cdot(3/4\cdot K_{3GPI}) \\
&+ TF_{BEC}(2\omega_1-\omega_2)\cdot(j(2\omega_1-\omega_2)\cdot 3/4\cdot K_{3CPI}) \\
&+ TF_{CBC}(2\omega_1-\omega_2)\cdot(j(2\omega_1-\omega_2)\cdot 3/4\cdot K_{3CBC} \\
&\cdot(TF(\omega_1)-1)^2\cdot(\overline{TF(\omega_2)}-1)) \\
&+ TF_{CEC}(2\omega_1-\omega_2)\cdot(1/4\cdot K_{3GM2GO}\cdot(2TF(\omega_1)+\overline{TF(\omega_2)})) \\
&+ TF_{CEC}(2\omega_1-\omega_2)\cdot(1/4\cdot K_{3GMGO2}\cdot TF(\omega_1) \\
&\cdot(TF(\omega_1)+2\overline{TF(\omega_2)}))\}\cdot V_{BE}(\omega_1)^2\cdot\overline{V_{BE}(\omega_2)}
\end{aligned}
\tag{4.24}
$$

The equations for IM3 caused by cascaded quadratic electrical nonlinearities become quite complicated, and a full set of equations, including 21 mixing products from the envelope and 21 from the second harmonic, is given in Appendix C. As an example, a double mixing caused by $K_{2GM}{}^{*}K_{2GPI}$ is explained here. This contribution is written as

$$
\begin{aligned}
&V_{CE9}(2\omega_1-\omega_2) = K_{2GM}\cdot K_{2GPI}\cdot[TF_{BEC}(2\omega_1-\omega_2) \\
&\cdot\overline{TF_{CEBE}(\omega_2-\omega_1)}+TF_{CEC}(2\omega_1-\omega_2)\cdot\overline{TF_{BEBE}(\omega_2-\omega_1)}] \\
&\cdot V_{BE}(\omega_1)^2\cdot\overline{V_{BE}(\omega_2)}
\end{aligned}
\tag{4.25}
$$

This equation includes two low-frequency mixing mechanisms, which are sketched graphically in Figure 4.9. In the first mechanism, illustrated in Figure 4.9(a), a product of $V_{BE}(\omega_1)$ and $V_{BE}(-\omega_2)$ in K_{2GM} generates the envelope current $\omega_1-\omega_2$ between the collector and the emitter, from where it is then transferred back to the base-emitter voltage by TF_{CEBE}. These envelope and fundamental voltages $V_{BE}(\omega_1-\omega_2)$ and $V_{BE}(\omega_1)$ are further mixed in K_{2GPI}, generating IM3 current at $2\omega_1-\omega_2$ between the base and the emitter. This is further converted to the collector voltage in the transfer function TF_{BEC}. The second mechanism, shown by Figure 4.9(b), can be explained as follows: $V_{BE}(\omega_1)$ and $V_{BE}(-\omega_2)$ generate the envelope current $\omega_1-\omega_2$ directly at the base in K_{2GPI}. This current between the base and the emitter is transferred to the base-emitter voltage by TF_{BEBE}, and the resulting envelope voltage mixes with the fundamental base-emitter voltage

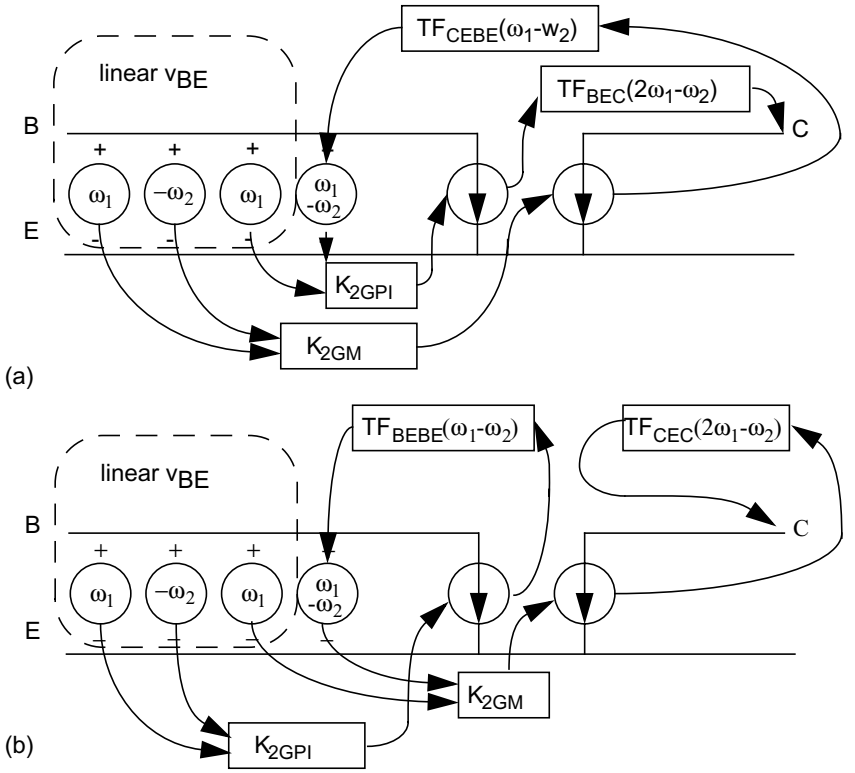


Figure 4.9 IM3L caused by the cascaded second-degree distortion mechanisms K_{2GPI} and K_{2GM} via the envelope frequency. From [11].

in K_{2GM}, as a result of which an IM3 current is generated between the collector and the emitter. This current is finally converted to the collector voltage by TF_{CEC}.

Such multiple mixing products are quite common. For example, one major cause of IM3 in deep class AB or class B amplifiers is the following: When clipping asymmetrically, the transistor causes a high second harmonic at the output (this is modeled by high K_{2GM}). As the frequency of the second harmonic is high, it couples easily through C_{BC} back to the input and mixes with the fundamental again in K_{2GM}, causing IM3 current directly in the output. This mechanism (named as term V_{CH1} in Appendix C) can be minimized by attenuating the second harmonic either at the collector or at the base.

4.3.2.5 Electrothermal Terms

Finally, the third-order electrothermal distortion mechanisms are described. These consist of third-degree terms $TF_{XYZ}*(K_{3XXT}*v*t_J)$ and cascaded second-degree terms, where the thermally induced second-order distortion mixes with fundamental tones in electrical square-law nonlinearities. The third-order terms can be expressed by

$$\begin{aligned} &V_{C3T3}(2\omega_1-\omega_2) \\ &= TF_{CEC}(2\omega_1-\omega_2)\cdot(K_{3GOT}\cdot TF(\omega_1)\cdot V_{BE}(\omega_1)\cdot T_J(\omega_1-\omega_2)) \\ &+ TF_{CEC}(2\omega_1-\omega_2)\cdot K_{3GMT}\cdot V_{BE}(\omega_1)\cdot T_J(\omega_1-\omega_2) \\ &+ TF_{BEC}(2\omega_1-\omega_2)\cdot K_{3GPIT}\cdot V_{BE}(\omega_1)\cdot T_J(\omega_1-\omega_2) \\ &+ j(2\omega_1-\omega_2)K_{3CPIT}\cdot TF_{BEC}(2\omega_1-\omega_2)\cdot V_{BE}(\omega_1)\cdot T_J(\omega_1-\omega_2) \\ &+ j(2\omega_1-\omega_2)K_{3CBCT}\cdot TF_{CBC}(2\omega_1-\omega_2) \\ &\cdot[TF(\omega_1)-1]\cdot V_{BE}(\omega_1)\cdot T_J(\omega_1-\omega_2) \end{aligned} \tag{4.26}$$

where T_J is the junction temperature, TF is the V_{CE}/V_{BE} ratio, and TF_{XYC} is the transimpedance from the nonlinear source to collector. Note that the pattern is always the same: low-frequency temperature variations modulate the value of the circuit element, which causes mixing with the fundamental tone. The junction temperature is calculated based on the power dissipation $P_{diss} = v_{CE}i_C$, which results in both first- and second-order tones. Only the envelope tone $P_{diss}(\omega_2-\omega_1) = (V_{CE}(\omega_2)I_C(-\omega_1) + V_{CE}(-\omega_1)I_C(\omega_2))/2$ is picked and multiplied by the thermal impedance $Z_{TH}(\omega_2-\omega_1)$ to obtain the variation in the junction temperature. The dissipating source is always the same, but if different temperatures are needed for different circuit elements, the thermal network can be modeled by a multiport Z-matrix.

The IM3 contributions of cascaded second-degree nonlinearities are again too complicated to be presented here, but these are shown in Appendix C, and their numerical values are plotted in the vector plots in Sections 4.4.2 and 4.5. Note that the terms including a temperature-dependent charge K_{2CXXT} cannot be calculated unless we find a way to measure or derive a value for it.

As seen from the above and Appendix C, already the symbolic Volterra analysis of a simple CE amplifier gets quite involved, and more complex topologies are too difficult to handle analytically. Volterra calculations can be performed numerically as well, however. The Nlsim software [24, 29] is capable of calculating the nonlinear voltage components numerically, and more complex topologies can be analyzed. Compared to the fully analytical solution, Nlsim gives a slightly less detailed picture of distortion, as it does not separate the cascaded quadratic nonlinearities but simply displays the cubic term and up- and downconverted envelope and second harmonic terms for each nonlinearity. However, the second-order phasors can be plotted as vector sums to see the dominant second-order contributions.

4.4 IM3 in a BJT CE Amplifier

Here, IM3 distortion is studied in two different cases: first using a simplified cascade model of nonlinear input impedance and I-V curve, and then the full circuit, including also the feedback effects, employing the full analysis presented in the previous section.

4.4.1 BJT as a Cascade of Two Nonlinear Blocks

The distortion composition of a BJT amplifier is quite complicated, as noted in the previous section, and therefore some simplifications are first made here to find some of the basic characteristics of distortion composition. A complete analysis will be given later, but for the moment the feedback effects of the emitter impedance Z_E and feedback capacitance C_{BC} are ignored and the collector current is simply considered to be one-dimensional, a function of base voltage only. The simplified circuit is shown in Figure 4.10 and since all feedback effects have been ignored, the collector current of the BJT can now be regarded simply as a cascade of two nonlinear blocks, as shown in Figure 3.5. The first block contains the input nonlinearities C_{pi} and g_{pi} and the second the nonlinearity of the one-dimensional transconductance.

It is commonly known that the linearity properties of the CE BJT amplifier are different when using either voltage or current excitation. In

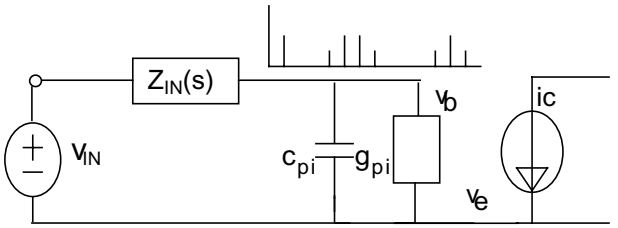


Figure 4.10 The simplified BJT model, ignoring the feedback effects. © IEEE 2000 [30].

the idealized case the voltage excitation with zero Z_{IN} makes the input nonlinearities negligible, because the input and base-emitter voltages are equal, and the voltage source shorts all distortion currents. Then the only effective nonlinear element in Figure 4.10 is the transconductance. The situation changes if a current source excitation is applied. Figure 4.11(a) shows the base-emitter voltage as a function of input current and Figure 4.11(b) the collector current as a function of the base-emitter voltage, assuming purely exponential g_{pi} and g_m. Figure 4.11(a) can be recognized as a logarithmic function, while Figure 4.11(b) is an exponential one, and as a cascade of them, the collector current as a function of input current is perfectly linear, as shown in Figure 4.11(c). Thus, the strong exponential nonlinearity of a voltage-driven BJT disappears, when the transistor is driven by a current. Also, in practice the type of the excitation can be adjusted somewhat by the value of Z_{IN}: the higher the value of Z_{IN}, the more the CE BJT appears to be current driven.

In Figure 4.11 the two opposite nonlinearities cancel one another. The same phenomenon is also demonstrated in the Gummel plot in Figure 4.12, in which the collector and base currents as functions of V_{BE} are presented. Since a logarithmic y-axis is used, these purely exponential nonlinearities appear as straight lines. It can be observed that the distance (i.e., the current gain) between the curves is constant and independent of the value of V_{BE}, indicating that the shapes of the nonlinearities are similar. These two nonlinearities canceling each other out are often referred to as tracking nonlinearities, resulting in linear operation with nonlinear elements. In practice, however, the nonlinearities do not track completely, as a result of which some amount of nonlinearity always exists. For example, the dashed parts of the lines in Figure 4.12 represent the practical situation in which the transconductance is not purely exponential at high V_{BE} values anymore due to a high injection level at the base, thus reducing the current gain. Also, at very low V_{BE} values some leakage base current always exists.

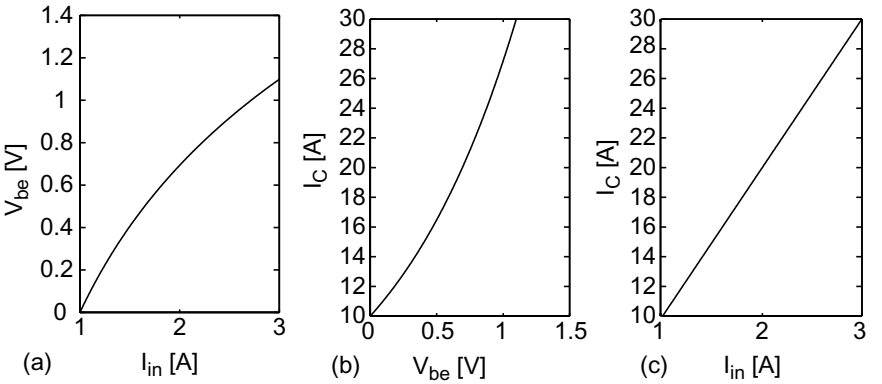


Figure 4.11 Illustration of tracking nonlinearities: (a) V_{be} as a function of I_{in}, (b) I_C as a function of V_{be}, and (c) Ic as a function of I_{in}.

Next, the ac behavior of the circuit shown in Figure 4.10 is analyzed. The nonlinearity of the first block is caused by g_{pi} and C_{pi}; g_{pi} is considered to be purely exponential, and C_{pi} is exponential, too, being calculated from the forward transit time (τ_F) and transconductance g_m as follows [31]:

$$C_{pi} = \tau_F \cdot g_m . \tag{4.27}$$

The latter block represents the nonlinear transconductance, and its nonlinearity can be calculated from (4.1). If we ignore the Early effect and use only one-dimensional collector current, the nonlinearity coefficients listed in Table 4.2 can be obtained. The second column corresponds to the strictly exponential low injection case, and the fourth column presents a high injection condition, in which the nonlinearity of the transconductance is reduced, corresponding to the dashed line in Figure 4.12.

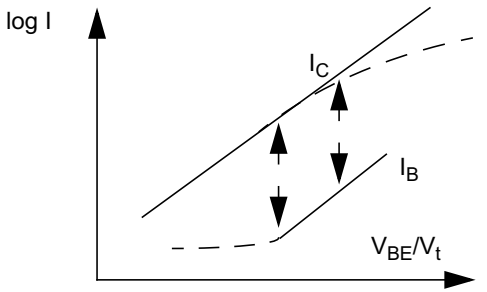


Figure 4.12 Base and collector current as a function of base-emitter voltage. Dashed parts of the curves present high injection and leakage current effects.

Table 4.2
Nonlinearity Coefficients for Transconductance (From [30])

Degree of nonlinearity	$I_C \ll I_{KF}$	$I_C = 10^{-2} \cdot I_{KF}$	$I_C = 10^{-1} \cdot I_{KF}$
1	I_C / V_T	$0.981 \cdot I_C / V_T$	$0.845 \cdot I_C / V_T$
2	$I_C / (2 \cdot V_T^2)$	$0.962 \cdot I_C / (2 \cdot V_T^2)$	$0.724 \cdot I_C / (2 \cdot V_T^2)$
3	$I_C / (6 \cdot V_T^3)$	$0.925 \cdot I_C / (6 \cdot V_T^3)$	$0.535 \cdot I_C / (6 \cdot V_T^3)$

The distortion composition of cascaded nonlinearities was discussed in Chapter 3 and is shown graphically in Figure 3.6. The IM3L collector current of the cascade is found to be

$$\begin{aligned} i_{\mathrm{OUT}}(2\omega_1-\omega_2) = {} & V_{\mathrm{BE}}(2\omega_1-\omega_2) \cdot g_{\mathrm{m}} \\ & + 3/4 \cdot V_{\mathrm{BE}}^2(\omega_1) \cdot \overline{V_{\mathrm{BE}}(\omega_2)} \cdot K_{3\mathrm{GM}} \\ & + \overline{V_{\mathrm{BE}}(\omega_2)} \cdot V_{\mathrm{BE}}(2\omega_1) \cdot K_{2\mathrm{GM}} \\ & + V_{\mathrm{BE}}(\omega_1) \cdot V_{\mathrm{BE}}(\omega_1-\omega_2) \cdot K_{2\mathrm{GM}} \end{aligned} . \tag{4.28}$$

IM3L consists then of four components, the first two of which are generated directly inside the first and the second block, respectively. The last two are related to second-order interaction via the envelope and second harmonic frequencies between the blocks. In the first block the distortion is generated by the second-degree nonlinearity of the input impedance and the spectral components at the base for (4.28) can be calculated using Table 2.6.

The effects of the fundamental Z_{IN} of the circuit in Figure 4.10 are studied first. The nonlinearity coefficients are calculated using equations (4.1) and (4.27), based on the GP model SPICE parameters taken from a BFG11 RF power BJT [32]. The IM3 levels are calculated for a center frequency of 1.8 GHz and tone spacing of 1 MHz. If the value of the input impedance Z_{IN} is high compared to the internal base impedance, the input will behave like a current source and the input-output nonlinearities will partially cancel each other out.

Z_{IN} at the fundamental is swept over the range of reasonable values to check the effects of voltage/current excitation. The input of the BJT is

conjugately matched, which means that the input reactance was chosen so that the base capacitance is tuned out (j*imag(Z_{IN}) = +j0.6). Z_{IN} values at the harmonic and envelope are set to infinity, which means that the base impedance at these frequencies consists of the internal base-emitter impedance only. IM3 distortion components at two values of the collector current are shown as functions of fundamental Z_{IN} in Figure 4.13. The IM3 asymmetry (i.e., the vector error between the lower and upper sidebands) is also given in Figure 4.13.

Under low injection conditions, IM3 improves by increasing the value of Z_{IN} up to 20 to 30 ohms. Increasing the value of Z_{IN} any further does not reduce IM3 any more, because incomplete tracking between the input and output nonlinearities limits IM3 to a level of –57 dBc, as seen in Figure 4.13. When approaching high injection conditions (I_C = 0.1*IKF*), the situation is different: although the linearity is quite independent of Z_{IN} at very high values of the latter, a significant linearity improvement can be achieved at some relatively small values. The higher I_C is compared to the knee current *IKF*, the better is the linearity improvement achievable at optimum Z_{IN}.

When the imaginary part of the fundamental Z_{IN} is increased from conditions of conjugate match, the linearity starts to deteriorate, but even though it may be reduced slightly, the vector difference (i.e., the

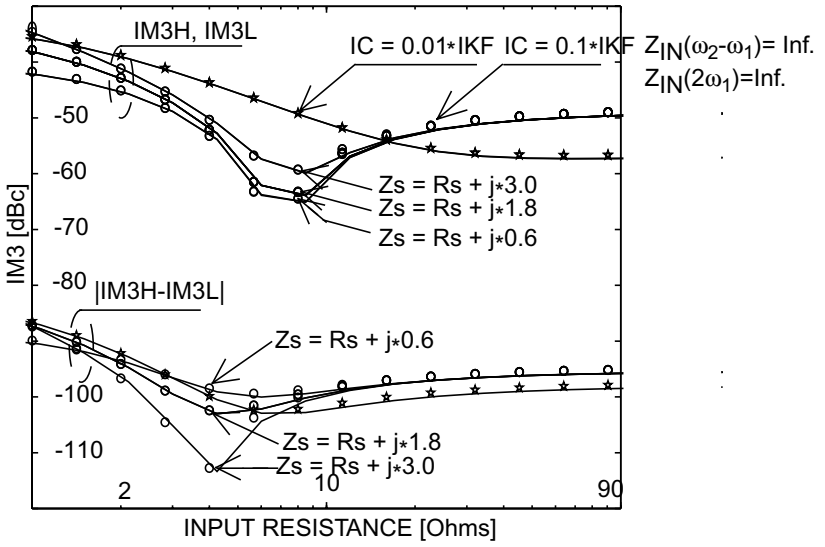


Figure 4.13 Amplitude of the IM3 sidebands of the collector current as a function of the fundamental input impedance. © IEEE 2000 [30].

asymmetry) between the IM3 sidebands may reduce. The lower curves in Figure 4.13 represent the asymmetry between the lower and upper IM3 tones. By making the input match slightly inductive, a significant improvement of more than 10 dB may be achieved in the symmetry of the IM3 sidebands. Similar observations have been reported in [33], where the effects of input impedance on phase distortion were studied. The improvement in the symmetry can be a very important advance in applications involving linearization.

Since a large amount of the output third-order distortion is generated via cascaded second-degree nonlinearities, careful optimization of out-of-band Z_{IN} can improve the linearity. Third-order terms cannot be affected much by filtering, because their power overlaps the fundamental signal, but the power of the second-order signal lies well away from the fundamental, and filtering can be used to achieve the best possible linearity [27, 34-36]. Also, there are quite a lot of requirements for the fundamental matching such as gain and efficiency, so that the fundamental impedance cannot be chosen based only on the linearity properties. However, the out-of-band matching has only secondary effects on gain and efficiency, and therefore it can sometimes be tuned mainly based on the linearity.

We will now look at the effects of second-order signal components. The impedance around the fundamental is kept constant while Z_{IN} at the envelope is swept. The upper eight-segment vector in Figure 4.14(a) represents the output IM3 product caused by input nonlinearities (C_{pi} and g_{pi}), while the two first parts of the lower three-segment vector represent the output third-order IM3 products caused by second-degree input-output nonlinearities. The first of these is generated via the envelope at the base and the other one via the second harmonic. The third part of the lower three-segment vector represents the cubic nonlinearity of the transconductance. It is interesting to note that the parts of the IM3 vector caused by the quadratic nonlinearities are opposite to the part caused by cubic nonlinearities. This means that the nonlinearities track each other in two different ways: First, through the input-output tracking explained earlier, which means that the distorted voltage waveform at the base cancels the distortion caused by nonlinear transconductance, and second, through the cancellation between the second- and third-degree nonlinearities. The latter may be difficult to note from the equations presented in this book due to extensive use of transfer functions. However, if we ignore C_{pi} and write the expression for the third harmonic current in terms of circuit elements and coefficients in (4.29) [2, equation 8.76], the opposite signs of the second- and third-degree coefficients are clearly visible.

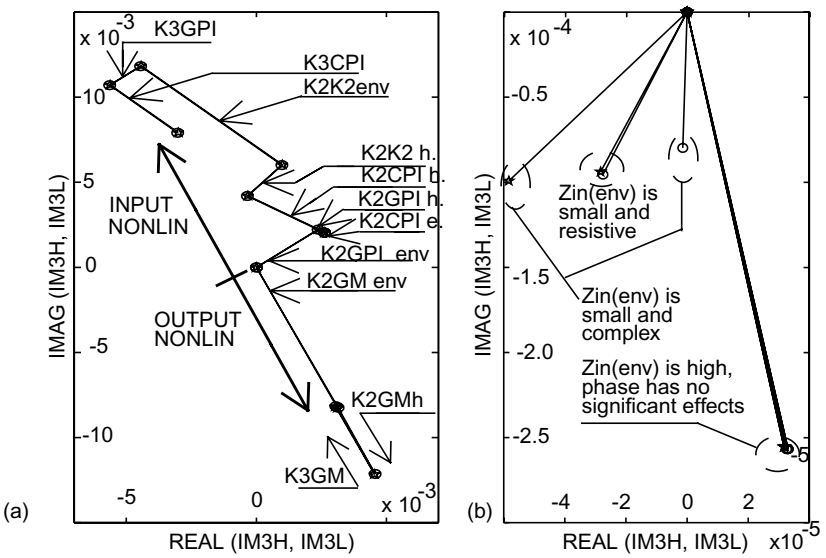


Figure 4.14 Representation of (a) partially tracking nonlinearities, and (b) the result of vectors. Both output IM3 signals consist of 11 contributors: K2GPIe, K2GPIh, K2CPIe, K2CPIh, K2K2GPICPIh, K2K2GPICPIe, K3GPI, K3CPI, K2GMe, K2GMh, and K3GM (e and h mean envelope and second harmonic, respectively). © IEEE 2000 [30].

$$I_{\mathrm{OUT}}(3\omega) = \frac{I_{\mathrm{in}}^3}{4g_{\mathrm{pi}}}(-g_{\mathrm{pi}} \cdot K_{3\mathrm{GM}} + g_{\mathrm{m}}g_{\mathrm{pi}}K_{3GPI} + 2g_{\mathrm{pi}}K_{2\mathrm{GM}}K_{2\mathrm{GPI}} - 2g_{\mathrm{m}}K_{2\mathrm{GPI}}) \tag{4.29}$$

The result of the IM3 caused by these partially tracking nonlinearities is illustrated in Figure 4.14(b) for four values of the impedance at the envelope. If the magnitude of that impedance is high, the output IM3 signal will not be affected by its phase, but if it is relatively small, say at most one decade higher than the Z_{IN} at the fundamental frequency, its phase will play an important role. The reason why Z_{IN} at the envelope has an impact on IM3 only when it is small is simple: If Z_{IN} is high, the total impedance of the base node is dominated by the internal impedance of the transistor, but when Z_{IN} is small, it affects the total node impedance.

Minimizing the low frequency Z_{IN} has often been suggested as a way to reduce overall distortion. As seen above, it is not always the case; instead, a small imaginary value for Z_{IN} may reduce the total distortion, or improve the balance between the IM3 sidebands. Since distortion currents multiplied by $Z_{IN}(\omega_2-\omega_1)$ and $Z_{IN}(\omega_1-\omega_2)$ mix to different sidebands and the impedances have opposite phases [as $Z_{IN}(\omega_1-\omega_2) = \overline{Z}_{IN}(\omega_2-\omega_1)$], a small and reactive baseband bias impedance may be quite handy in correcting phase errors between the upper and lower IM3 sidebands.

4.4.2 Detailed BJT Analysis

Now the full Volterra model is employed to study the effects of individual distortion mechanisms. In this case the nonlinearity coefficients of the model are extracted by measurements, using a procedure that will be explained in detail in Chapter 5. In this section, the effects of the bias point and optimum terminal impedances at different frequencies on the linearity of the CE BJT amplifier are discussed.

It is a well-known fact that the optimum impedances of a CE amplifier vary with the biasing conditions. It is impractical to examine experimentally all possible combinations of fundamental input and load impedances at different bias voltages, and if the envelope and harmonic impedances, which have a great impact on linearity, were to be taken into account, too, optimization of a single CE stage would become a very lengthy task. The Volterra model – provided it is accurate enough – can be used for optimizing the linearity of a CE amplifier by designing optimal input and load impedances under different sets of biasing conditions.

An amplifier based on a BFG11 transistor [32] is constructed and the input impedance at a fundamental frequency of 1.8 GHz is tuned to the conjugate match for maximum power transfer. The load impedance is chosen so that the imaginary parts of the output reactance and matching network canceled each other out, and the real part of the Z_L is determined by the desired I-V characteristics. Linearity can be improved by lowering the value of R_L, because the voltage swing at the collector decreases, but unfortunately, the efficiency decreases at the same time. So R_L was chosen to be 20 ohms as a trade-off between linearity and efficiency. The load impedance around zero frequency is small, and has to be so to supply the dc energy from the supply to the collector with minimum losses. The input and load impedances of the amplifier are $0.1+j76$ and $31+j17$ ohms, respectively, at the second harmonic and $13-j0.2$ and $0.2-j1.4$ at the envelope frequency of 2 MHz. These figures are taken from measurements performed on the actual, implemented amplifier.

The base and collector bias voltages are first swept and the linearity monitored to find some basic characteristics of the biasing. To make the simulations comparable to each other, the input voltage swing and load resistance are tuned a little to keep the output voltage and current swings constant under varying biasing conditions. An output voltage swing of 1.5 Vpp is chosen, which means that the IM3 values near the collector supply of as low as 1V are not accurate. Calculated IM3L contours and measured IM3L points, in dBc, are shown in Figure 4.15(a), and a good correlation between these two is observed.

Maximum linearity is achieved at a bias current of 75 mA, and linearity deteriorates with a decreasing bias current. The IM3 sideband increases by approximately 10 dB when the bias current is lowered to 40 mA. The collector bias voltage affects the linearity as well, so that 3.5V gives the maximum linearity. Collector voltages beyond 4V are not shown in Figure 4.15(a), because higher collector voltages only reduce the efficiency without giving further linearity improvement. At low currents, the linearity is quite independent of collector voltage, but at high current values the collector voltage plays a more important role. This is obvious because at

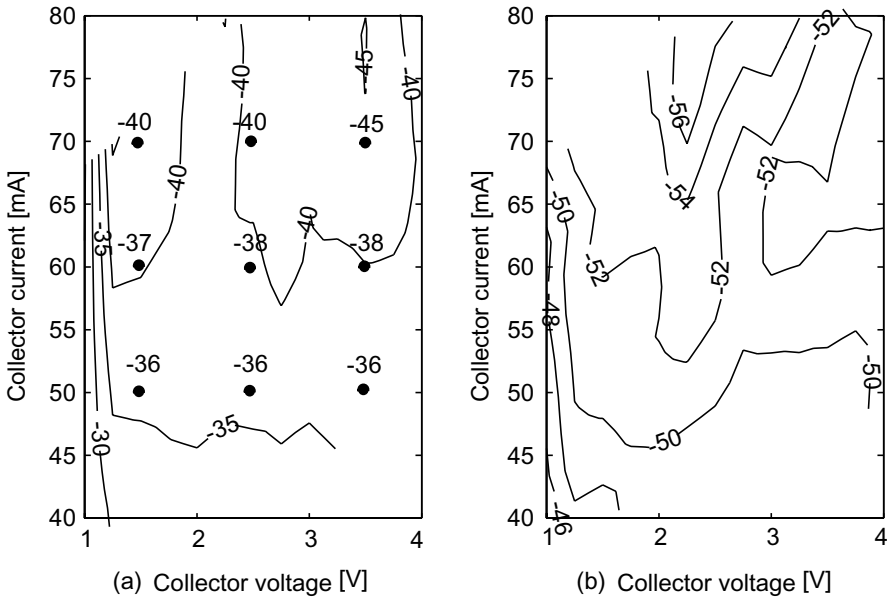


Figure 4.15 (a) Calculated and measured (dots) IM3L and (b) calculated asymmetry IM3L-IM3H as functions of the biasing condition for an output voltage swing of 1.5 Vpp. From [11].

lower collector current values the nonlinearity caused by g_{pi}, C_{pi}, and g_m is much stronger than that caused by g_o and cross-terms. Moreover, at high current values the nonlinearity caused by g_{pi}, C_{pi}, and g_m is weaker, because transconductance, for example, linearizes with an increasing current due to the effects of high injection. Also, since g_o and its nonlinearity are quite strong at high collector current values, g_o takes a more dominant role with respect to total distortion.

The contour plot for the IM3H sideband is quite similar to that for the IM3L in Figure 4.15(a). The vector difference between the sidebands is plotted in dBc compared to the fundamental signal in Figure 4.15(b). It is important to emphasize that this figure is closely dependent on modulation frequency and not only on bias values and matching impedances. The asymmetry between the sidebands is an important figure of merit if predistortion linearization is employed, as discussed in Chapter 3.

The values for the fundamental input and load impedances in the previous example are chosen to achieve the desired power and gain characteristics with reasonable efficiency. For the best possible linearity, however, the out-of-band envelope and harmonic terminations have to be optimized. One commonly used approach for improving the linearity by tuning the out-of-band terminations is to minimize the load impedance at the second harmonics. Since the IM3 components are partially caused by the second harmonic voltages, the impedance at that frequency affects IM3. Whether or not this improves the linearity depends on the nonlinearity coefficients and other impedances. To check the effects of the second harmonic matching, the real part of the load impedance at that frequency is reduced to 6 ohms. The result of the comparison is plotted in Figure 4.16, which shows improvements of 1 dB to 5 dB under all biasing conditions. These calculations show that optimization of out-of-band impedances is needed to achieve the best performance. Only the effects of the second harmonic load impedances were demonstrated, but other out-of-band effects are important, too. Carefully selected optimum out-of-band terminating impedances can improve the linearity by several decibels without reducing the power and efficiency performance at the fundamental.

The asymmetry between the IM3 sidebands is greatly affected by the envelope impedances, as noted in Chapter 3. In multicarrier transmitters the bandwidth of the signal may be very wide, and it is very hard to design constant impedances between dc and 20 MHz, for example. The input and load envelope impedances are usually not constant, and since the IM3 sidebands are functions of these impedances, they vary as a function of the modulation frequency, causing memory effects.

The memory effects are investigated by sweeping the tone spacing of a two-tone signal at collector and base bias voltages of 3V and 740 mV,

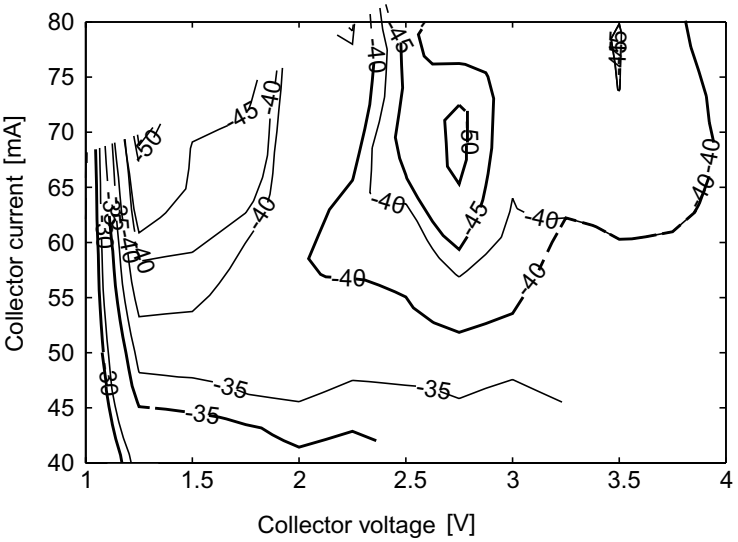


Figure 4.16 Calculated IM3L at different load impedances at the second harmonic (thin lines 35 ohms, thick lines 6 ohms). From [11].

respectively. The phase of the Volterra-calculated IM3L is plotted in Figure 4.17 as a function of tone spacing, and three different types of memory effects are met: thermal memory effects at low frequencies, resonance-type memory effect, and high-frequency memory effects, both caused by the biasing circuits. The resonance at 1 MHz is caused by a resonating collector impedance. It cannot be observed with harmonic balance (HB) simulations using the Gummel-Poon (GP) model, because its oversimplified output impedance masks the effect of the collector resonance. The smooth phase deviation at high modulation frequencies is caused by the input impedance at the envelope frequency, and it is nevertheless simulated correctly by the GP model, too. Since the dynamic self-heating effects are not implemented in the basic GP model, it naturally cannot predict the low frequency memory effects caused by thermal effects. The Volterra simulations agree reasonably well with the measured results, given also in Figure 4.17 and explained in more detail in Chapter 6. The Volterra model therefore seems to be a good tool for recognizing memory effects in a power amplifier.

Next we will look at the fine structure of the IM3 phasors. This information can be used in design optimization, and IM3 is drawn here as a vector sum of tens of contributions to study why the phase of the IM3L

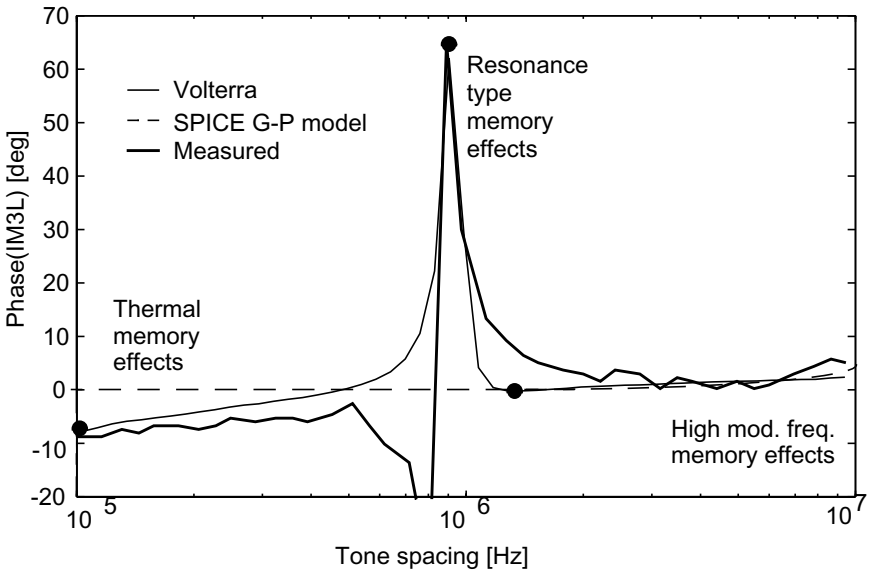


Figure 4.17 Calculated, simulated, and measured phases of total IM3L as a function of tone spacing. From [11].

depends on modulation frequency. Three frequencies are chosen for the vector plots, 100 kHz, 1 MHz, and 1.1 MHz, and the resonance at 1 MHz and the thermal memory effects are studied.

We first look what happens just above the resonance at tone spacing of 1.1 MHz. The IM3L phasor is plotted in Figure 4.18(a) as a vector sum on a real-imaginary axis, starting from the origin at the top center. The vector consists of 49 purely electrical terms, the first seven of which are caused by cubic nonlinearities and the following 42 are generated by cascaded second-degree nonlinearities via the envelope and second harmonic frequencies. The electrothermal terms are practically zero, because the 1.1 MHz beat frequency already lies in the stopband of the thermal filter.

The first seven points (1-7) are the cubic distortion mechanisms, of which $K_{3\mathrm{CPI}}$ (6) is the largest. The following 21 (8-28) are upconverted IM3 components from the envelope frequency, and finally, the last 21 (29-49) are downconverted from the second harmonic. One interesting finding is that there is not a single dominant contribution that we can attack; instead, the total IM3 is already smaller than any of the dominant contributions. This is due to several pairs that partially cancel each other:

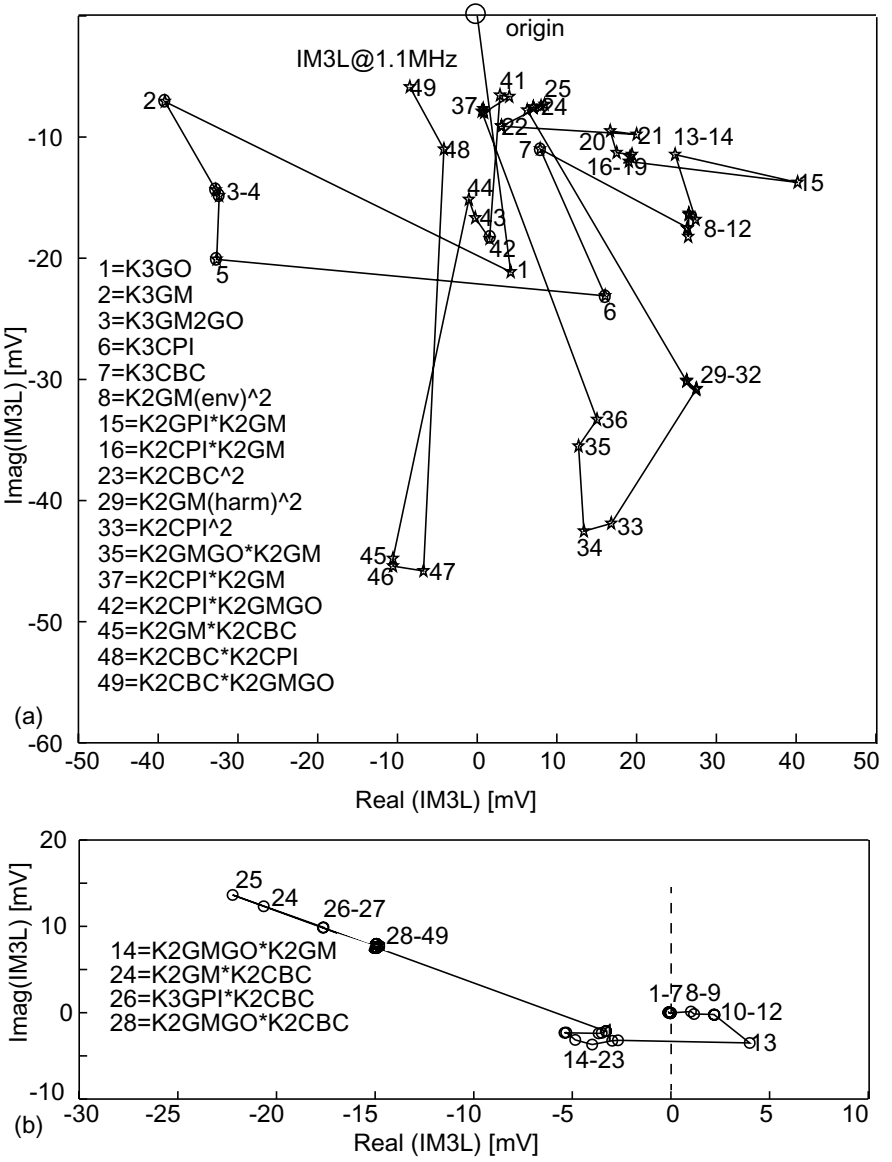


Figure 4.18 Vector representations of (a) IM3L@1.1MHz, and (b) the vector difference between IM3L@1.1MHz and IM3L@1.0MHz. From [11].

the third-degree nonlinearities K_{3GM} (2) and K_{3CPI} (6) get partially canceled, and so do the second harmonic mixing results K_{2GM}*K_{2GM} (29, where the second harmonic caused by g_m returns to the input and mixes again), and K_{2CPI}*K_{2GM} (37, second harmonic caused by C_{pi} mixing in K_{2GM}), or K_{2GM}*K_{2CBC} (45, second harmonic caused by g_m mixing in C_{BC}) and K_{2CBC}*K_{2CPI} (48, second harmonic caused by C_{pi} being amplified and mixed in C_{BC}).

What could be done to improve linearity? Distortion proportional to K_{3GPI} is converted to voltage in fundamental Z_{IN}, and its contribution (6) can be rotated towards the origin by adjusting the phase of Z_{IN}. Further, the size of the entire 29-49 mesh can be reduced by lowering the second harmonic impedance. On the other hand, a 5% to 10% increase of the base impedance at the second harmonic would increase terms 37 and 48 and force the total sum closer to zero.

From the memory effect point of view it is instructive to study what happens when the tone spacing is reduced to 1 MHz, hitting exactly the resonance in Z_L. This is illustrated in Figure 4.18(b), which depicts the difference of IM3L@1.1MHz and IM3L@1.0MHz as a similar vector sum. First, the cubic nonlinearities (1)-(7) behave in exactly the same way, as the phases of the fundamental tones do not vary. The second harmonic mixing products (29-49) are also equal, indicating flat terminal impedances at the second harmonic. All of the dominant causes of the IM3 resonance arise from cascaded quadratic nonlinearities that are upconverting the envelope frequency - large terms include K_{2GMGO}*K_{2GM} (14, output envelope mixing in the K_{2GMGO}), K_{2GM}*K_{2CBC} (24, output envelope mixing in C_{BC}), and K_{2GPI}*K_{2CBC} (26, the input envelope being amplified and mixed in C_{BC}), that all involve the baseband frequency response of Z_L.

Memory effects generated by interaction between electrical and thermal behavior are studied next at a narrow tone spacing of 100 kHz. Purely electrical IM3 vectors are presented by the first segments of the vectors in Figure 4.19, and at such a low frequency they are almost identical. The electrothermal 30-segment vector is drawn in a termwise manner, and the electrothermal distortion is seen to be dominated by large K_{3GMT}, K_{2GMT}, K_{3CPIT}, and K_{2CPIT} terms, which after all cancellations at the end point 78 cause only a 15% difference to IM3 amplitudes caused by purely electrical mechanisms. Note, however, the big phase difference between IM3L and IM3H components. It is again due to the fact that baseband effects, including the thermal feedback, mix with opposite phases to lower and higher IM3 sidebands. Besides phase asymmetry, the thermal feedback also causes here amplitude asymmetry between IM3L and IM3H, as the distance between the origin and the final point 78 is different for IM3L and IM3H.

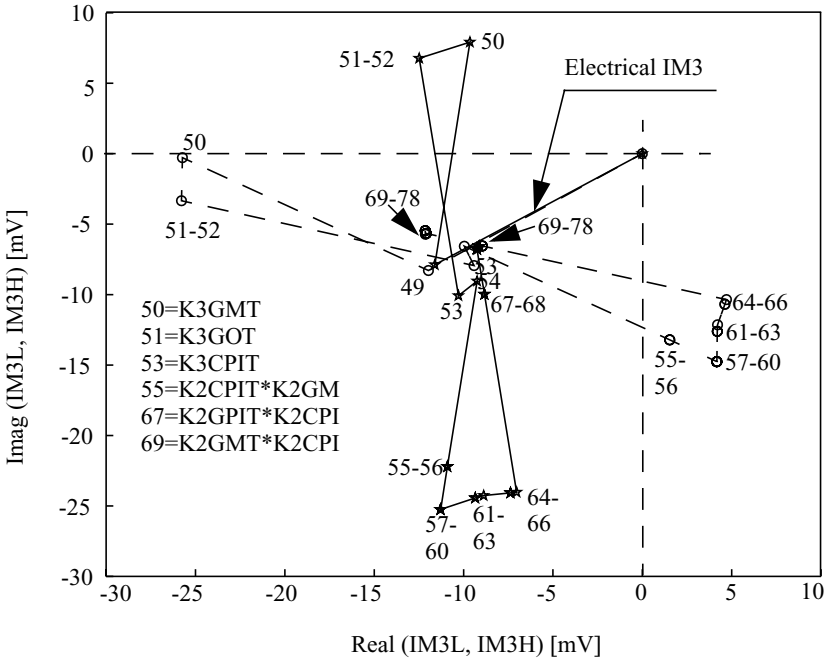


Figure 4.19 Vector representation of electrothermal IM3L (dashed) and IM3H (solid) distortion. Tone spacing is only 100 kHz to see the thermal effects. From [11].

4.5 MESFET Model and Analysis

A Volterra model for a MESFET common source amplifier is derived in this section. Its small-signal equivalent circuit, given in Figure 4.20, includes an input impedance Z_{IN} (consisting of the driver impedance, external input impedance, and internal series gate impedance), a gate-to-source capacitance (C_{GS}), a feedback capacitance (C_{GD}), a drain-to-source capacitance (C_{DS}) and a resistance ($1/g_o$), transconductance (g_m), a load impedance (Z_L), and a source impedance (Z_S). As in the BJT case, the input and load impedances include not only impedances of the matching networks, but also impedances of the bias networks. The drain current of the FET is modeled as a three-dimensional function of the gate and drain voltages and temperature similar to (4.3) for a BJT. C_{GS} and C_{GD} are also

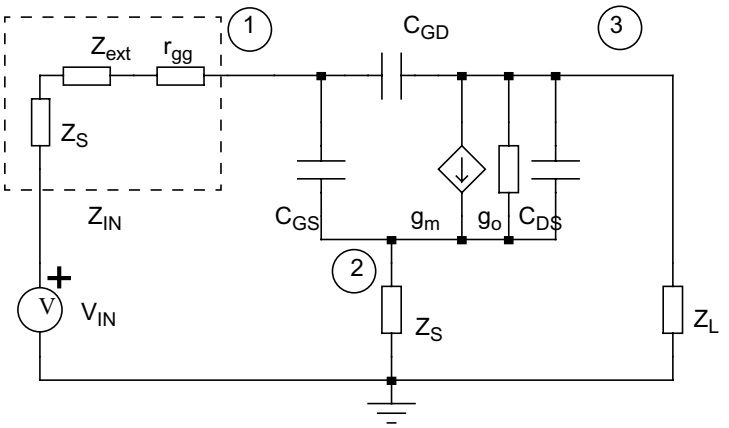


Figure 4.20 Linearized first-order circuit for a common-source FET amplifier.

regarded as nonlinear, being functions of the gate-to-source voltage and temperature and of the drain-to-gate voltage and temperature, respectively.

The linearized circuit for a common-source FET amplifier presented in Figure 4.20 is pretty much the same as that for a BJT amplifier shown in Figure 4.7. By neglecting the g_{pi} and changing the names of the circuit elements, the equations deduced in Section 4.3.2 can be used for the FET analysis as well. Similarly, by omitting g_{pi} and its nonlinear current source, a circuit containing a distortion source for an FET can be obtained, as given in Figure 4.21. The nonlinearities of the device are characterized by 22 first-, second-, and third-degree nonlinearity coefficients, of which 15 are purely electrical (C_{GS}, K_{2CGS}, K_{3CGS}, C_{GD}, K_{2CGD}, K_{3CGD}, g_m, K_{2GM}, K_{3GM}, g_o, K_{2GO}, K_{3GO}, K_{2GMGO}, K_{3GM2GO}, and K_{3GMGO2}) and seven are related to temperature variations (K_{2CGST}, K_{3CGST}, K_{2CGDT}, K_{3CGDT}, K_{2GMT}, K_{3GMT}, and K_{3GOT}). Equations (4.3), (4.6), (4.24), (4.26), and (4.28) as well as Appendix C can be used for an FET simply by ignoring g_{pi} and changing the voltages and subscripts BE to GS and BC to GD.

Circuit elements and nonlinearity coefficients for an Infineon CLY2 GaAs MESFET [37] are extracted using an *S*-parameter characterization method discussed in Chapter 5. The input and load impedances measured from an implemented amplifier are listed in Table 4.3.

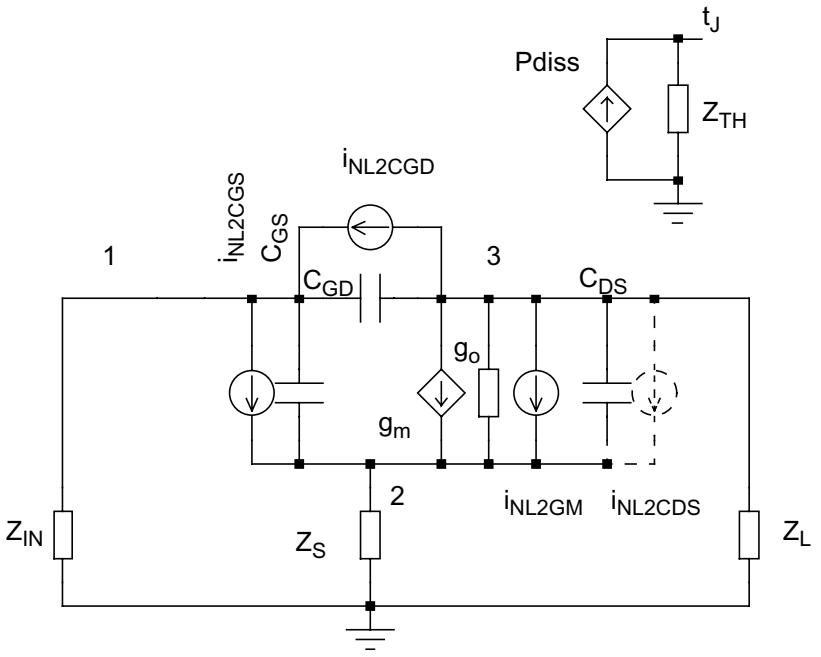


Figure 4.21 Circuit containing second-order distortion sources. The thermal ones can be obtained by connecting $i_{NL2CGST}$ in parallel with i_{NL2CGS} and i_{NL2GMT} with i_{NL2GM}.

Table 4.3
Input and Output Matching Impedances at Different Frequency Bands

Frequency band	Z_{IN}	Z_L
2 MHz (envelope)	$1.3k - j1.3k$	$0.51 - j4.2$
1.8 GHz (fundamental)	$6.6 + j11$	$26 - j6.6$
3.6 GHz (2nd harmonic)	$18 - j47$	$3.6 - j8.7$

The IM3 calculations over the range of bias current and voltage at the center and modulation frequencies of 1.8 GHz and 2 MHz and with an output voltage swing of 2 Vpp are depicted by a solid line in Figure 4.22. Large linearity variations of more than 30 dB are observed over the I-V

plane, but since the dc power consumption changes with bias point according to the linearity, no significant advance in terms of the efficiency-linearity trade-off can be achieved. The Volterra calculations agree reasonably well with the measured linearity values represented by the dots in Figure 4.22.

Next, the effects of the out-of-band impedances are studied. The value of the input impedance at the envelope frequency is increased from 1.3k–*j*1.3k to 10k – *j*10k to see the effect of bias impedances. The new linearity contours drawn with dashed lines in Figure 4.22 show some significant changes. The linearity is reduced by a few decibels at low drain voltages, but very good linearity improvements of more than 10 dB are observed at high voltages of 4V and a drain current value of 150 mA. Since Z_{GG}(env) is highly frequency-dependent, however, the improvements presented in Figure 4.22 are quite narrowband.

The changes caused by the modified gate impedance at the envelope frequency are demonstrated in Figure 4.23, where IM3 vectors at 2-MHz

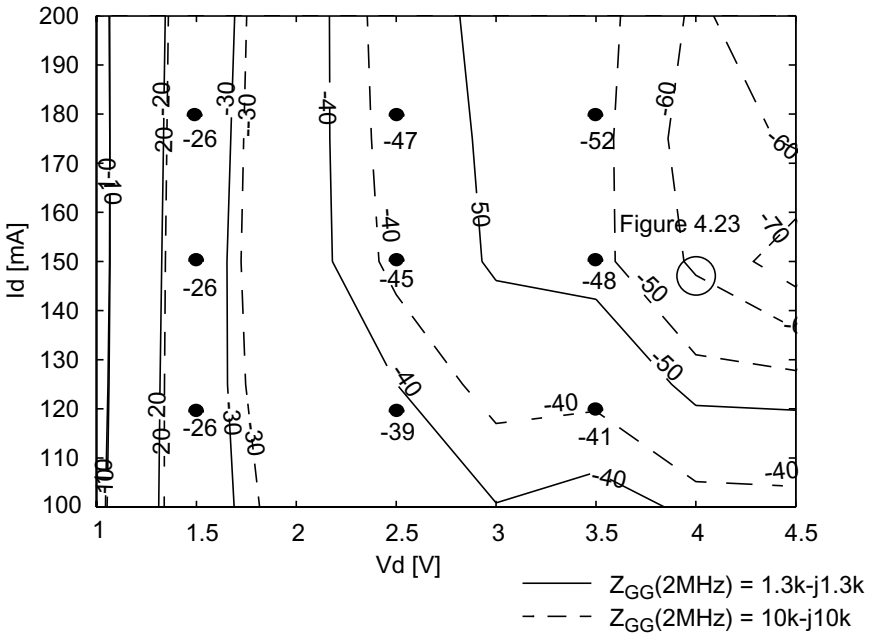


Figure 4.22 Calculated (lines) and measured (dots) IM3L contours in dBc at two values of the input impedance at the envelope frequency. Center and modulation frequencies are 1.8 GHz and 2 MHz, and output voltage swing is 2 Vpp. © IEEE 2002 [38].

tone spacing for the two baseband input impedances are calculated at the bias point of V_D=4.5V and I_D=150 mA, shown in Figure 4.22. The total distortion is dominated by the cubic distortion mechanisms K_{3GM}, K_{3GM2GO}, K_{3GMGO2}, and K_{3CGS}, and some cancellation between the mechanisms can be seen in the figure. The effect of the cross-terms, especially K_{3GM2GO}, is very significant, and most of the large contributions are due to the I-V characteristic. However, the reason for improvement at the higher Z_{GG} is caused by C_{GS} (term K_{2CGS}^2, shown for Z_{GG}=10k–*j*10k only): The larger gate impedance amplifies the envelope current generated in K_{2CGS}, and after it mixes again in K_{2CGS} to IM3, it mostly cancels the large K_{3CGS} term, as shown in Figure 4.23.

The effects of gate impedance at the envelope frequency have been investigated also in [39], where linearity changes caused by varying the gate impedance have been demonstrated by measurements. The same effect has been measured for the baseband drain impedance in [40, 41]. Optimum gate and drain impedances were found to be low in all of these cases [39-41], but it is shown here that the optimum envelope impedance can be either high or low, depending on the amplitudes and phase angles of the nonlinearity mechanisms. An optimum nonzero envelope impedance was

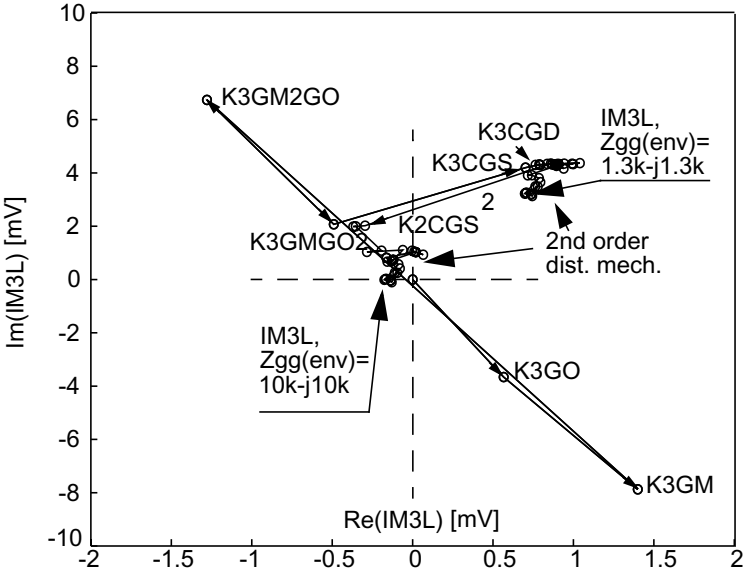


Figure 4.23 Vector representations of IM3 components for two gate bias impedances. The bias point is Vd=4V, Id=150mV and the tone spacing is 2 MHz. © IEEE 2002 [38].

also found by measurements in [35]. In conclusion, the effects of out-of-band impedances can be used for optimizing the distortion behavior, and a certain amount of improvement in IM3 performance can be expected. In general the effects of out-of-band impedances are often undesirable, however, because the envelope impedance (mostly caused by the dc bias circuit) cannot be constant over a wide range of modulation frequencies. Since the IM3 components are affected by envelope impedances, memory effects will occur.

To study the memory effects, a tone spacing sweep is again simulated using the Volterra model. The phases of the IM3L and IM3H tones, plotted in Figure 4.24, show a very good agreement with the measured results (presented in more detail in Chapter 6). The bump at 500-kHz tone spacing is also in this case caused by a resonance in the envelope drain impedance, and the phase drift at very high tone spacings is caused by a frequency-dependent baseband gate impedance. This 5º to 10º variation in the phase is certainly not serious in a standalone amplifier, but it is enough to reduce the achievable cancellation below 15 dB in a predistorted PA, as will be seen in Chapter 6.

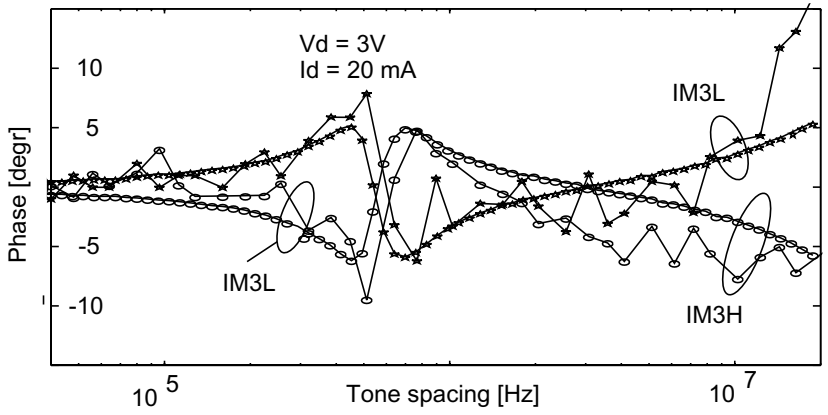


Figure 4.24 Calculated (smooth) and measured (rough) phases of the IM3 components as functions of tone spacing. © IEEE 2002 [38].

4.6 Summary

The requirements for the simulation models are challenging and since all the requirements cannot be fulfilled, the models are optimized for different purposes. Most of the models for transistors/amplifiers can be divided into two classes: behavioral and device models. Behavioral models just try to imitate the measured phenomena without any information on internal device operation, while device models more or less imitate the physical operation of the device. A lot of research is going on in developing the device models, but still they are not completely optimized for RF power amplifier design. Instead, most semiconductor models are aimed for analog IC design, where a wide range of operating conditions are used, and models need to be scalable for different sizes and geometries. From an RF power amplifier design point of view, accurate distortion simulations are essential, and the derivatives of the I-V and Q-V curves and self-heating effects should be as accurate as possible. Unfortunately, only a few manufacturers provide parameters for the most sophisticated power transistor models.

The Volterra analysis is an extension to small-signal analysis, and the nonlinearities of the circuit elements are modeled by polynomial functions that are extracted around the desired bias voltages. In this way component-level information is achieved and the model is fitted locally just around the chosen operating point, thereby providing good accuracy in simulations. The main advantage of the model and using the Volterra analysis for distortion simulation is that elementwise information of IM3 can be obtained: IM3 can be drawn as a sum of vectors, each presenting the nonlinearity of one circuit element or mixing mechanism. In this way the Volterra analysis provides insight into distortion mechanisms, and gives information about the dominant ones and possible cancellation schemes. This is very helpful when optimizing the matching networks and selecting proper bias voltages.

The third-order Volterra model presented in this chapter can be used for both BJT/HBT and MESFET common emitter/source amplifiers, with slight modifications. The model includes a nonlinear three-dimensional collector/drain current which is a function of the base-emitter/gate-source and collector-emitter/drain-source voltages and of temperature. The input capacitance C_{pi}/C_{GS} is a nonlinear function of the base-emitter/gate-source voltage and temperature. The feedback capacitance C_{BC}/C_{GD} is also a nonlinear function of the collector-base/drain-gate voltage and temperature. In a BJT model, g_{pi} is a function of the base-emitter voltage and temperature, although its effect is small whenever the transistor is used at high frequencies, where C_{pi} dominates the input distortion. Since the impact of the other circuit elements on total distortion is small, these are

considered to be linear components. The model presented in this chapter also includes the feedback capacitance C_{BC}/C_{GD} and the emitter/source impedance Z_E/Z_S, which have been neglected in most previously published Volterra models. The circuit is solved analytically, and complete equations for the IM3 components are presented here and in Appendix C.

The collector/drain current used in this analysis is a complete Taylor expansion of its variables, and compared to most other Volterra models, the input-output cross-products K_{2GMGO}, K_{3GM2GO}, and K_{3GMGO2}, which have a significant impact on distortion performance, are taken into account. The temperature on the surface of the chip is proportional to the instantaneous power dissipation $v_{CE}i_C$. Thus the junction temperature is already a second-order phenomenon, and it is modeled as an independent, low frequency variable. The thermally induced IM3 can be calculated using the Volterra model, and it affects IM3 at low modulation frequencies of up to hundreds of kilohertz.

The IM3 at the output is not only the sum of the effects of cubic nonlinearities, but also the cascaded quadratic nonlinearities have an impact on the total amount of distortion. Therefore, the out-of-band impedances at the envelope $\omega_2-\omega_1$ and second harmonic $2\omega_1$ can be used to optimize the distortion behavior, and linearity improvements of some decibels can be achieved by using optimum out-of-band impedances. Since the impedances, and especially the envelope impedance, cannot be constant over a wide range of modulation frequencies, the amplitude and/or phase of IM3 becomes dependent on the modulation frequency, which is very harmful with many linearization techniques. These memory effects can be simulated with the Volterra model, which is capable of predicting the measured memory effects with sufficient accuracy.

The impact of out-of-band impedances and electrical memory effects seem to be stronger in BJTs than MESFETs. In BJTs, there are two strong, almost exponential nonlinearity mechanisms that partially cancel out each other. This phenomenon is very sensitive not only to fundamental but also to out-of-band impedances, and IM3 can be significantly affected by the latter. In both BJT and MESFET, collector/drain impedance at the envelope frequency is difficult to design, because large LC time constants are needed for energy storage, causing frequency-dependent envelope impedance and hence memory effects. These effects are more serious in BJTs, but special attention also has to be paid to designing the baseband Z_{GG} in MESFETs. Due to C_{GS}, this impedance is high and markedly tilted, and some amount of memory effect is generated at high modulation frequencies.

Dynamic thermal effects and TPF are more important in BJTs than in MESFETs. The thermal impedances of the chips and packages are quite similar, but due to the fact that the electrical circuit elements in a BJT are

more sensitive to temperature, more TPF is generated in BJTs. In most cases, TPF has to be taken care of in a BJT, whereas in a MESFET the phase of the IM3 starts to be affected by dynamic temperature variations on the surface of the chip only when the electrical IM3 value is already very small.

The total IM3 consists of a number of distortion mechanisms, and many of them partially cancel out each other. This tracking phenomenon is dependent on matching impedances and nonlinearity coefficients, which in turn are dependent on the bias voltages. It is often observed in practice that linearity suddenly improves at some value of the bias voltages, for example. This is caused by canceling nonlinearities, and in some situations the tracking is very good and good linearity is achieved. Unfortunately, tracking is very sensitive to changes in impedances, voltages, and temperature, which makes it difficult to exploit experimentally. The Volterra model presented in this chapter nevertheless provides a systematic way for studying these effects, so that it is easy to see whether or not tracking is possible and to which parameters the tracking is sensitive. A careful investigation into distortion mechanisms is the key to successful exploitation of the internal cancellation mechanisms.

4.7 Key Points to Remember

1. Accurate distortion simulations are needed in RF power amplifier design.
2. In general, simulation models can be divided into two classes: behavioral and device models. Device models may be based on predefined functions, or just on tabulated measured data.
3. Accurate derivatives have traditionally been only a secondary requirement in developing new device models.
4. The derivatives up to the order of N of the I-V and Q-V curves must be accurate enough for Nth-order distortion simulations.
5. Using the polynomial Volterra model, the derivatives can be locally fitted into the actual behavior.
6. The Volterra analysis can give – either analytically or numerically – the response of each distortion mechanism separately. This makes it possible to look at the total distortion in a termwise manner that gives a lot of information for design optimization.

7. There are many partially canceling distortion mechanisms in RF power amplifiers. For example, the g_{pi} and g_m nonlinearities cancel each other in a current-driven BJT.
8. PA designers have very limited control to the intrinsic nonlinearities of the device, but they can affect the transfer functions that convert the distortion currents generated by the nonlinearities to node voltages.
9. The IM3 caused by cascaded quadratic nonlinearities is visible in both BJT and MESFET amplifiers, especially in BJTs.
10. TPF is more important in BJT/HBT than MESFET amplifiers, because the circuit elements of the BJT are more temperature-dependent.
11. The Volterra model is a powerful tool for recognizing the memory effects of the amplifier.
12. Low-frequency bias impedances cause most of the electrical memory effects.

References

[1] Maas, S., "How to model intermodulation distortion," *1991 IEEE MTT-S International Microwave Symposium Digest*, Vol. 1, pp. 149-151.

[2] Wambacq, P., and W. Sansen, *Distortion Analysis of Analog Integrated Circuits,* Norwell, MA: Kluwer, 1998.

[3] Webster, D., et al., "Effect of model derivative discontinuities on cold FET distortion simulations," *1997 Workshop on High Performance Electron Devices for Microwave and Optoelectronic Applications*, pp. 97-102.

[4] Tsividis, Y., and K. Suyama, "MOSFET modeling for analog circuit CAD: problems and prospects," *IEEE Journal of Solid-State Circuits*, Vol. 29, No. 3, 1994, pp. 210-216.

[5] Root, D., "Principles and procedures for successful large-signal measurement-based FET modeling for power amplifier design," *Gain Without Pain seminar material*, Agilent Technologies, 2000.

[6] Le Gallou, N., et al., "An improvement behavioral modeling technique for high power amplifiers with memory," *2001 IEEE International Microwave Symposium*, Phoenix, AZ.

[7] Kenington, P.B., *High Linearity RF Amplifier Design,* Norwood, MA: Artech House, 2000.

[8] Chen, J., "What the nonlinear K-model does and how to make it do more," Cadence Design Systems, 2000, at http://www.cadence.com/datasheets/rf_notes.html

[9] Verbeyst, F., and V. Bossche, "VIOMAP, the S-parameter equivalent for weakly nonlinear RF microwave devices," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 42, No. 12, 1994, pp. 2531-2535.

[10] Wang, T., and T. Brazil, "A Volterra mapping-based S-parameter behavioral model for nonlinear RF and microwave circuits and systems," *1999 IEEE MTT-S International Microwave Symposium Diges*t, Vol. 2, pp. 783-786.

[11] Vuolevi, J., "Analysis, measurement and cancellation of the bandwidth and amplitude dependence of intermodulation distortion in RF power amplifiers," Doctoral thesis, University of Oulu, Oulu, Finland, 2001.

[12] McAndrew, C., and L. Nagel, "Early effect modelling in SPICE," *IEEE Journal of Solid-State Circuits*, Vol. 31, No. 1, 1996, pp. 136-138.

[13] *HSPICE User's Manual Release 96.1*, Meta-Software Inc., 1996.

[14] Laker, K., and W. Sansen, *Design of Analog Integrated Circuits and Systems*, New York: McGraw-Hill, 1994.

[15] Versleijen, M., and A. Bauvin, "Accuracy of bipolar compact models under RF power operating conditions," *Proc. 1994 IEEE MTT-S International Microwave Symposium Digest*, Vol. 3, pp. 1583-1586.

[16] De Graaff, H., et al., "Experience with the new compact MEXTRAM model for bipolar transistors," *Proc. 1989 Bipolar Circuits and Technology Meeting*, pp. 246-249.

[17] De Vreede, N., et al., "Advanced modeling of distortion effects in bipolar transistors using the Mextram model," *IEEE Journal of Solid-State Circuits*, Vol. 31, No. 1, 1996, pp. 114-121.

[18] Van Rijs, F., et al., "RF power large signal modeling with MEXTRAM," *Proc. 1996 Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 57-60.

[19] Kloosterman, W., J. Geelen, and D. Klaassen, "Efficient parameter extraction for the MEXTRAM model," *Proc. 1995 Bipolar/BiCMOS Circuits and Technology Meeting*, pp. 70-73.

[20] McAndrew, C., et al., "VBIC95, the vertical bipolar inter-company model," *IEEE Journal of Solid-State Circuits*, Vol. 31, No. 10, 1996, pp. 1476-1483.

[21] Curtice, W., et al., "A new dynamic electro-thermal nonlinear model for silicon RF LDMOS FETs," *1999 IEEE MTT-S International Microwave Symposium Diges*t, Vol. 2, pp. 419-422.

[22] *Microwave Office^{TM} User's Manual II*, Applied Wave Research, Inc., 2000.

[23] Kolding, T., and T. Larsen, "High order Volterra series analysis using parallel computing," *International Journal of Circuit Theory and Application*s, Vol. 25, No. 2, 1997, pp. 107-114.

[24] Heiskanen, A., and T. Rahkonen, "5th order multi-tone Volterra simulator with component-level output," *Proc. 2002 IEEE International Symposium on Circuits and Systems*, Phoenix, AZ, 2002, pp. 591-594.

[25] Bin, L., and S. Prasad, "Intermodulation analysis of the collector-up InGaAs/InAlAs/InP HBT using Volterra series," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 46, No. 9, 1998, pp. 1321-1323.

[26] Lee, J., et al., "Intermodulation mechanism and linearization of AlGaAs/GaAs HBT's," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 45, No. 12, 1997, pp. 2065-2072.

[27] Fong, K., and R. Meyer, "High-frequency nonlinearity analysis of common-emitter and differential-pair transconductance stages," *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 4, 1998, pp. 548-555.

[28] Crosmun, A., and S. Maas, "Minimization of intermodulation distortion in GaAs MESFET small-signal amplifiers," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 37, No. 9, 1989, pp. 1411-1417.

[29] Rahkonen, T., and J. Vuolevi, "Term-wise Volterra analysis of nonlinear analog circuits for recognising dominant cause of distortion," *Proc. Norchip 2000 Conference*, Turku, Finland, November 6-7, 2000, pp. 198-203.

[30] Vuolevi, J., and T. Rahkonen, "The effects of source impedance on the linearity of BJT common-emitter amplifiers," *Proc. 2000 IEEE International Symposium on Circuits and Systems*, Geneva, Switzerland, 2000, pp. IV-197-IV-200.

[31] Quarles, T. et al., *SPICE3 Version 3f3 User's Manual*, University of California, Berkeley, CA, 1993.

[32] *BGF 11/X NPN 2 GHz RF Power Transistor Datasheet*, Philips Semiconductors 1995

[33] Yamada, H., et al., "Self-linearizing technique for L-band HBT power amplifier: effect of source impedance on phase distortion," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 44, No. 12,2 1996, pp. 2398-2402.

[34] Aparin, V., and C. Persico, "Effect of out-of-band terminations on intermodulation distortion in common-emitter circuits," *1999 IEEE MTT-S International Microwave Symposium Digest*, Vol. 3, pp. 977-980.

[35] Sevic, J., K. Burger, and M. Steer, "A novel envelope-termination load-pull method for ACPR optimization of RF/microwave power amplifiers," *1998 IEEE MTT-S International Microwave Symposium Digest*, Vol. 2, pp. 723-726.

[36] Staudinger, J., "The importance of sub-harmonic frequency terminations in modelling spectral regrowth from CW AM-AM & AM-PM derived non-linearities," *Proc. 1997 IEEE Wireless Communications Conference*, pp. 121-125.

[37] *CLY 2 GaAs Power MESFET Datasheet*, Infineon Technologies, 1996.

[38] Vuolevi, J., and T. Rahkonen, "Extraction of nonlinear AC FET model using small-signal S parameters," *IEEE Trans. on Microwave Theory and Measurements*, Vol. 50, No. 5, May 2002, pp. 1311-1315.

[39] Tarui, Y., et al., "An improvement of IM and power of high power amplifiers using RC-paralleled circuits with frequency selectivity," *1998 IEEE MTT-S International Microwave Symposium Digest*, Vol. 3, pp. 1655-1658.

[40] Takenaka, I., et al., "Improved IMD characteristics in L/S-band GaAs FET power amplifiers by lowering drain bias circuit impedance," *IEICE Transactions on Electron*, Vol. 82, No. 5, 1999, pp. 730-736.

[41] Le Gallou, N., et al., "Analysis of low frequency memory and influence on solid state HPA intermodulation characteristics," *Proc. 2001 IEEE International Microwave Symposium*, Phoenix, AZ.

Chapter 5

Characterization of Volterra Models

In this chapter we will see how the parameters of Volterra models for RF power transistors can be extracted from measured data. Various indirect techniques have been used to build Volterra models, and for example the nonlinearities of g_m and g_o [1], and even the input-output cross-terms [2] have been estimated from the measured level of the harmonics. Some time-domain characterization methods have also been published [3, 4]. However, figures like harmonic distortion lump the effects of several nonlinearities, and we would like to build separate electrothermal models for each nonlinear I-V and Q-V source. Hence, the methods used here are quite similar to the techniques used in the extraction of any function based nonlinear device models. The main differences compared to conventional small-signal device characterization are:

- Power devices suffer from serious self-heating, and changing the drain bias voltages affects the power dissipation and hence the junction temperature. As the bias and the temperature vary simultaneously, these effects are difficult to separate. To overcome this problem, the self-heating can be minimized by arranging pulsed (isothermal) measurements.
- Volterra models are fitted locally around the desired bias point, and there is no need to model the entire range of bias voltages. Instead, the fitting range can be chosen according to the expected signal swing. Note also that the numerical sensitivity to small measurement errors can be affected by the location of the measurement points.
- Semiconductor manufacturers usually measure unpackaged dies. If we are characterizing packaged devices, de-embedding techniques are needed to remove the effect of the package from the measurements.

This chapter starts with a review of polynomial fitting techniques in Section 5.1. Then, the effects of self-heating and pulsed measuring techniques are presented in Section 5.2. I-V nonlinearities can be characterized by dc current measurements, and the required measurement setups and fitting techniques are presented in Section 5.3.

Q-V nonlinearities must be fitted using measured capacitance values, as charge cannot be measured using the normal measuring instruments. The ac measurements are usually made at RF frequencies and include several important topics that are discussed from Section 5.4 onwards. In more detail, pulsed *S*-parameter measurements and fixture calibration techniques are discussed in Section 5.5, removing the effects of a package from the measured data in Section 5.6, extracting the circuit element values from the measured *Y*-parameters in Section 5.7, and finally, fitting Volterra models based on the dQ/dV and dI/dV data obtained from the ac measurement in Section 5.8.

As examples, the measured and fitted results of three power transistors are presented. The extracted models for a 1W BJT and MESFET and a 30-W LDMOS device are discussed in Sections 5.9, 5.10, and 5.11, respectively.

5.1 Fitting Polynomial Models

5.1.1 Exact and LMSE Fitting

In the Volterra analysis, a nonlinearity $f(\mathrm{x})$ is presented as a series expansion (5.1) around the desired operating point $\mathrm{x_o}$:

$$f(x) = f(x_o) + a_1 \cdot (x - x_o) + a_2 \cdot (x - x_o)^2 + \ldots \tag{5.1}$$

where

$$a_k = \frac{1}{k!} \cdot \frac{\partial^k}{\partial x^k} f(x) \bigg|_{x = x_o} \tag{5.2}$$

However, it is by no means necessary to find the coefficients a_0-a_N by calculating the higher derivatives. Instead, normal polynomial fitting can be performed to directly find the coefficients. If we have exactly K equations, the fitting results will be exact at each measured point. However, the degree

of the polynomial should not be too high compared to the actual nonlinearity to be modeled. This is especially true if the original data is noisy, because errors in data points can cause slight oscillations in the fitting function between the data points. To avoid this, it is safe to use as low a degree polynomial as possible to model the nonlinearity accurately enough and always check the fitted result visually. If we have more data points, a least-mean-square error (LMSE) fitting can be performed. This result may not fit exactly to any of the measured points, but it is usually less sensitive to small errors in the measured data.

Polynomial fitting can be easily described using matrix operations. To fit a Kth-degree model using N results measured at points $x_1,x_2,...,x_N$, we group the measured results $f(x_i)$ into a Nx1 matrix $\boldsymbol{Y}$ and describe the model as a Nx(K+1) matrix $\boldsymbol{M}$ containing the different powers of x_i^j, i=1...N and j=0...K, calculated at each measurement point i, and a (K+1)x1 coefficient vector $\boldsymbol{A}$ containing the coefficients a_0 to a_K that need to be solved. Now the matrix equation (5.3) describes the system to be solved:

$$\boldsymbol{M} \cdot \boldsymbol{A} = \boldsymbol{Y} \tag{5.3}$$

where

$$\boldsymbol{M} = \begin{bmatrix} 1 & x_1 & .. & x_1^K \\ 1 & x_2 & .. & x_2^K \\ : & : & & : \\ 1 & x_N & .. & x_N^K \end{bmatrix} \qquad \boldsymbol{A} = \begin{bmatrix} a_0 \\ a_1 \\ : \\ a_K \end{bmatrix} \qquad \boldsymbol{Y} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ : \\ f(x_N) \end{bmatrix} \tag{5.4}$$

If now N=K+1, an exact solution for a_0-a_K is obtained simply by

$$\boldsymbol{A} = \boldsymbol{M}^{-1} \cdot \boldsymbol{Y}. \tag{5.5}$$

If the number of data points is larger than K+1, a LMSE fit is achieved by

$$\boldsymbol{A} = (\boldsymbol{M}^T \cdot \boldsymbol{M})^{-1} \cdot (\boldsymbol{M}^T \cdot \boldsymbol{Y}) \tag{5.6}$$

The contents of the matrix $\boldsymbol{M}$ is not limited to a one-dimensional polynomial. Instead, it may contain the selected powers and cross-products of V_{GS}, V_{DS} and temperature, for example. However, the measurement

points x_i must be chosen so that the matrix $\boldsymbol{M}$ will not become ill-conditioned, because the relative error of solution $\boldsymbol{A}$ is the relative error of measurements $\boldsymbol{Y}$ multiplied by a term proportional to the condition number of the model matrix $\boldsymbol{M}$.

Note that the Volterra analysis considers ac responses only. Hence, all voltages used in the $\boldsymbol{M}$ matrix are ac values, that is, $v = V - V_Q$, where V_Q is the bias point where the fit is performed. The fitted function $f(x)$ may consist of large signal values like currents or charges (if measurable), or be obtained from small-signal measurements, like capacitances or conductances.

5.1.2 Effects of Fitting Range

As the correct bias point must be obtained before the polynomial fitting can be applied, also the fitting range has to be correct. The limit values for the fitting range can be summarized as follows:

1. Large enough to encompass the nonlinearities;
2. Small enough to avoid effects lying outside the signal swing;
3. The ratio between the ranges must be correct in the case of multidimensional fitting.

The first requirement is obvious, because nonlinearities are difficult to detect if too narrow a range is selected, and a larger range will reduce the sensitivity to numerical errors. But, if the range is too large, nonlinear effects outside the signal swing will start to affect the extracted coefficients, even if they have no impact on the electrical distortion. This is illustrated in Figure 5.1(a), where the thick line represents the load line set by optimum R_L, the ellipse a dynamic load line, and the rectangular box the fitting range. The nonlinearities of the transistor are important only within this range, but if the nonlinearities are extracted over a larger range, nonlinear effects such as saturation will cause errors in the extracted coefficients and lead to erroneous or less accurate simulations.

The third requirement, illustrated in Figure 5.1(b), arises from two or more dimensional nonlinearities. The load resistance describes the slope of the load line corresponding to the ratio between the voltage and current swings. However, points cannot be chosen only from the load line, as this results in an insolvable group of equations. If all drain current measurements are chosen from a single load line so that $v_{DS} = -A_v v_{GS}$, the model functions like v_{GS}^3 and v_{DS}^3 become linearly dependent (as $v_{DS}^3 = -A_V^3 v_{GS}^3$) and the matrix equation (5.3) will be insolvable.

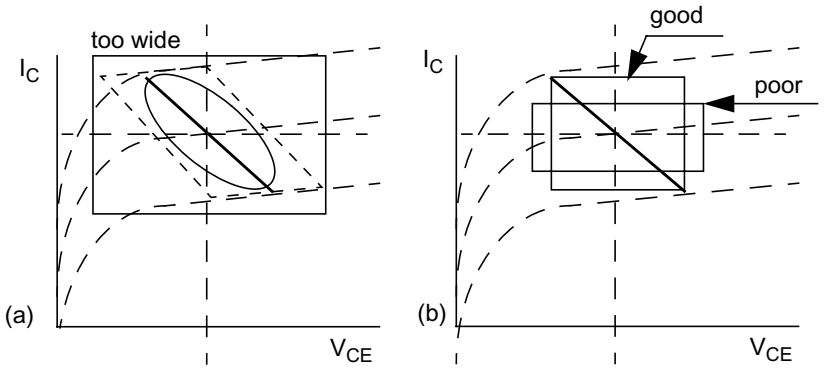


Figure 5.1 (a) A fitting range that is too wide in relation to the load line, and (b) good and poor fitting ranges in multidimensional fitting. From [5].

Hence, it is necessary to scatter the measurements to a wider area than just one line. However, if the drain voltage fitting range is too wide and the gate voltage too small, as drafted in Figure 5.1(b), some nonlinear effects of the output conductance outside the actual signal swing will be taken into account, but all nonlinear effects of the transconductance inside the actual signal swing will not be seen. The fitting range in both directions must correspond to the extent of the signal swing, as indicated in Figure 5.1(b).

As a conclusion, the fitting range should equal the actual signal swing in all dimensions. From a practical point of view, these requirements for a I_C-V_{BE}-V_{CE} curve can be fulfilled by two parameters, output power and load resistance. Signal swings in different directions can be calculated from these two parameters, and throughout this book, a box limited by the input and output voltage swings is used as a fitting area. This is assumed to provide a good approximation whenever the nonlinearities are so weak that they can be modeled by the third- or fifth-degree polynomials used here. Alternatively, a tilted area following the estimated load line [drawn with dashed line in Figure 5.1(a)] can also be used.

5.2 Self-Heating Effects

Nonlinear extraction is impeded by variations in chip temperature. Because the junction temperature responds to changes in biasing conditions, the circuit elements of the transistor model at different bias points will be extracted at different chip temperatures. The chip temperature of the biased and zero-input device can be calculated by

$$T_J = T_A + R_{TH} \cdot V_D \cdot I_D, \tag{5.7}$$

where T_A is the ambient temperature and R_{TH} is the thermal resistance (thermal impedance at dc) describing the temperature rise caused by dissipated power. In the characterization measurements, the power of the measured signal is small so it can be assumed that the dissipated power equals $V_D I_D$. As a result of (5.7), the constant chip temperature contours presented in Figure 5.2 are obtained. The temperature variations as a function of the bias values cause problems to the extraction of electrical and thermal nonlinearity coefficients. Let us illustrate this by rewriting (4.3) using just the two terms below:

$$i_D = g_o \cdot v_D + K_{3GOT} \cdot v_D \cdot t_J . \tag{5.8}$$

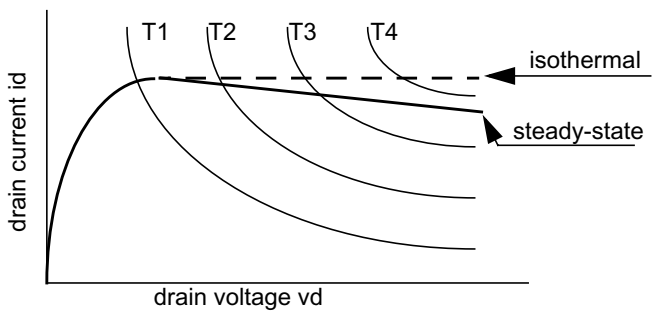


Figure 5.2 Constant temperature contours and extracted I-V curve with a CW (solid) or pulsed (dashed) signal source. From [5].

For simplicity, let us further consider g_o to be zero and K_{3GOT} to be negative. Now isothermal measurements yield the dashed line shown in Figure 5.2. This is a horizontal line with respect to v_D, corresponding to a zero g_o, as expected. However, if CW measurements are used, the solid line with a negative slope is obtained. This is due to the negative K_{3GOT}, which apparently seems to decrease the value of g_o. Actually this effect is due to changes in the chip temperature (proportional to $v_D i_D$), and if the temperature effects are not taken into account during the extraction, g_o wrongly derives a negative value. In other words, the value of the nonlinear output conductance changes with respect to both v_D and temperature, and

in CW measurements, these two effects are difficult, although not completely impossible, to separate from each other. Similar problems are encountered also with the other model elements, not just with the output conductance g_o.

Separation of electrical and thermal effects is required, because in power amplifier applications the drain voltage varies at the RF frequency, but the temperature changes only a little at the modulation frequency. Therefore, the terminal voltages and the junction temperature have to be independent variables, as discussed in Chapter 4. It was noticed that it would be possible to derive the terms for (5.8) using CW measurements also, if R_{TH} is accurately known. In some works the problems of self-heating are handled by measuring the transistor in a steady-state condition and then mathematically separating the effects of self-heating from the purely electrical behavior [6, 7]. Unfortunately, this is not possible in most power transistors because the dc power consumption can cause thermal breakdown at high bias values, destroying the device. This is the other reason why pulsed measurements are commonly used.

5.2.1 Pulsed Measurements

To avoid self-heating, the device must either be actively cooled, or, more simply, the dc bias must be pulsed with a low duty cycle to keep the average temperature constant [8, 9]. The effects of the width and duty cycle of the biasing pulse in pulsed measurements will be discussed next.

First of all, the measuring equipment must be fast enough to be able to measure during the pulsed bias. The pulse must be long enough to produce an electrical steady state, while at the same time it must be as short as possible to keep self-heating insignificant.

The range of the thermal time constants is illustrated with two examples. First, the thermal time constants of a packaged 1-W CLY2 MESFET transistor is measured by triggering the gate voltage to zero (on) and then monitoring the drain current as a function of time at ambient temperature of 20ºC with an oscilloscope. Any drift in the drain current indicates changes in the chip temperature, and thermal settling times of up to 3 seconds are seen in Figure 5.3. If the package is immersed in water (modeling a perfect heat sink), the largest visible settling time is 100 ms. In both cases, 1 ms can be considered a good measurement time, because 80% to 95% of the change in I_{DS} occur after this time. A similar order of magnitude of results is obtained in [10] by simulations. The optimum pulse length is dependent on the transistor type and package, so that the result cannot be generalized. To be sure that the self-heating can be neglected without considering the type, size, or transition frequency of the transistor,

the pulse should be shorter than the 1 ms used here. However, very fast measurements may cause problems with electrical settling. Thus, the measurement time should be as short as possible, but without causing electrical settling problems.

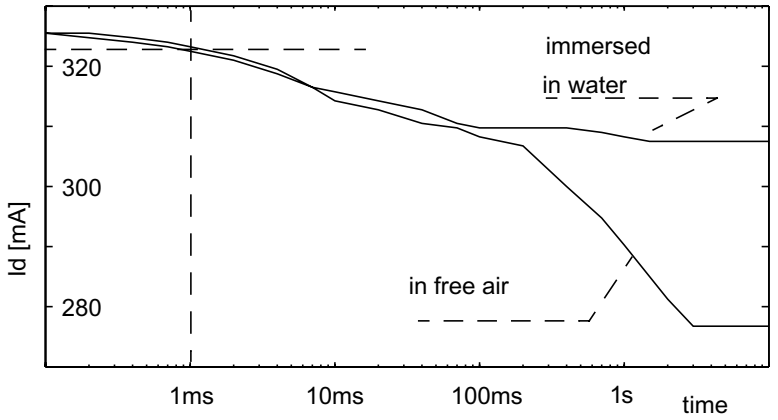


Figure 5.3 Thermal step response of a CLY2 chip and package after triggering on the gate bias. © IEEE 2002 [11].

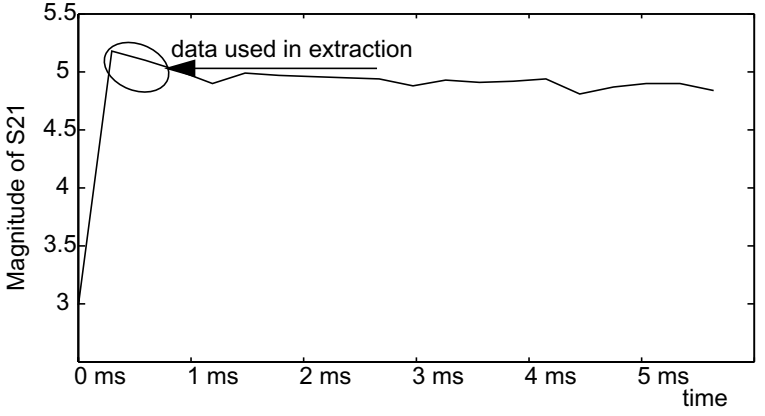


Figure 5.4 Small-signal S_{21} of an LDMOS transistor as a function of time after triggering on the bias voltage. © IEEE 2002 [12].

As another example, Figure 5.4 shows the magnitude of small-signal ac gain S_{21} of a 30W LDMOS MRF21030 as a function of time after turning the device on. This figure is obtained using the RF test setup presented in Section 5.5. The time resolution of the measurement is 333 μs, and it can be seen that the electrical steady state is already obtained in the first measurement after triggering. Due to self-heating and negative dg_m/dT, the absolute value of S_{21} starts to decrease. Eventually, also a thermal equilibrium is obtained, and S_{21} stops drifting. However, the time scale in Figure 5.4 is too short to show the thermal steady state of S_{21}, but the decrease of it as a function of time due to self-heating can easily be seen. In general, the optimum time of measurement is a trade-off between electrical settling (dominated by ac couplings and the transient response of the power supplies) and self-heating, but in this case the optimum measurement point is limited by the 3000 measurements-per-second rate of the network analyzer used. In other words, measurement accuracy can still be increased using shorter on pulses and faster measuring equipment.

5.2.2 Thermal Operating Point

To capture up to the third-degree electrothermal nonlinearities, characterization measurements must be made at two different temperatures T_1 and T_2, assuming that the time-varying junction temperature, caused by the power dissipation $v_{CE}*i_C$, itself is a second-order phenomenon. Neither of the measurement temperatures usually equals the actual operating temperature, and we need to see how the Volterra model fitted using these extreme temperatures applies at some intermediate temperature T_j.

As an example, let us consider a third-degree nonlinear electrothermal conductance model shown in (5.9). This model should now be fitted for junction temperature T_j using the data measured at temperatures T_1 and T_2, as illustrated in Figure 5.5.

$$i = g \cdot v + K_{2G} \cdot v^2 + K_{3G} \cdot v^3 + K_{2GT} \cdot t_J + K_{3GT} \cdot t_J \cdot v \qquad (5.9)$$

However, we need to be careful if the amount of electrical nonlinearity varies with the operating temperature. The temperature terms in (5.9) are meant for calculating the electrothermal distortion, but in principle they could be used to correct the drift in the electrical coefficients as well. Unfortunately, the third-degree term $K_{3GT}*v*t_J$ corrects temperature drift in the linear term g only (imagine the terms reordered into the form $(g+K_{3GT}*t_J)*v$), and the model as such predicts the same amount of electrical second- and third-degree nonlinearity at all temperatures. To be

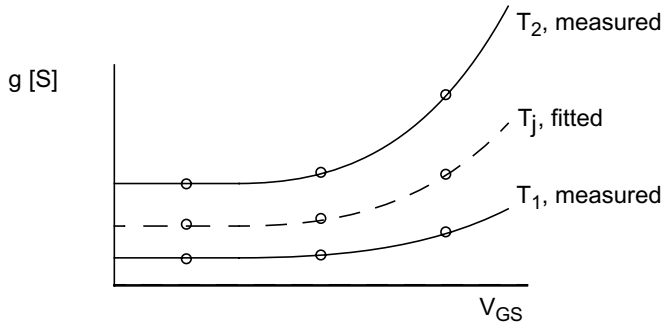


Figure 5.5 The interpolation of the small-signal elements correspond to the actual chip temperature.

able to model the temperature-dependent amount of electrical distortion, we need to do either one of the following two things:

1. The degree of the model must be increased. Addition of terms $K_{4GT}*v^2*t_J$ and $K_{5GT}*v^3*t_J$ allows the modeling of temperature-dependent v^2 and v^3 nonlinearities. In this case, a dc temperature term T_j-T_1 can be used to correct the drift in all nonlinear coefficients, and all parameters can be fitted simultaneously using directly the data measured at T_1 and T_2.
2. Alternatively, and more simply, the electrical nonlinearities can be fitted at the correct junction temperature. If data at T_j is not available, it can be obtained by taking the measured g,V pairs in both temperatures T_1 and T_2, interpolating a new $g(T_j),V$ set of data and fitting (5.9) to this data set. In this approach, data is interpolated before fitting.

The latter approach is used in the examples presented; thus, the data fitted is not necessarily the original data but an interpolated data set, corresponding to the operating temperature T_j between the measured temperature extremes.

Then what is the actual operating temperature? The Volterra analysis cannot give a solution to this, but it must be estimated somehow. One estimate can be formed using

$$T_J = T_A + R_{TH} \cdot V_D \cdot I_D \cdot (1 - \eta). \tag{5.10}$$

where η is the efficiency of the amplifier, R_{TH} is the total dc thermal resistance (including heat sinks and cooling), and V_D and I_D are the actual large-signal dc bias point. Unfortunately, Volterra analysis as described in Chapter 2 cannot calculate signal-induced shift in the bias point in one pass, but we have to either iterate or rely on measured bias point values. Moreover, as the efficiency depends strongly on the signal level, we usually have to fit a separate set of nonlinearity coefficients for each power level.

5.3 DC I-V Characterization

Pulsed dc measurements are often used for device characterization, as reported in [13, 14]. The main advantage of the dc I-V characterization is that the measurement setup is quite simple, and no high-frequency calibration is required. All the temperature drift terms can be characterized, which is not possible with the ac characterization technique described later. However, the degree of the fitted polynomial is higher than in ac characterization, due to the fact that the dc value also needs to be fitted. Thus, the dc fitting may be slightly more sensitive to numerical errors.

5.3.1 Pulsed DC Measurement Setup

Pulsed dc measurements are most easily arranged by applying a chosen dc voltage to the collector and then switching the base bias voltage on, as shown in Figure 5.6(a). Now the collector current is switched on, and a waveform similar to Figure 5.6(b) can be monitored using either a current probe or a small current-sensing resistor at the collector. The time required for achieving electrical steady state depends on the transient response of the collector voltage supply and any capacitors/inductors connected to the transistors, while the thermal settling typically takes up to several seconds. Isothermal collector current measurements can be made after the electrical settling but before the chip has warmed up due to self-heating. In the example presented in Section 5.9, the current is recorded with an oscilloscope 10 ms after triggering on the base voltage.

The V_{BE}-I_C control of a BJT is very steep and the control of the base voltage supply may not be accurate enough. In Figure 5.6(a) an external base series resistor is used to reduce the exponential nonlinearity of the transconductance and to make the transistor to behave more like current-driven. For correct modeling we now need to measure also the actual base voltage, and to model the I_B-V_{BE} nonlinearity, we also need to record the base current. If available, the multichannel oscilloscope and power supplies can of course be replaced with a pulsed high-power curve tracer.

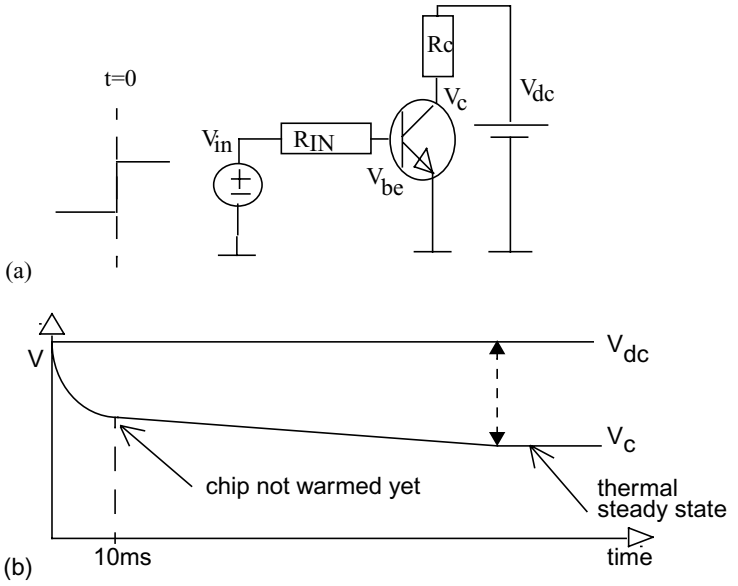


Figure 5.6 (a) A test bench for pulsed dc measurements, and (b) the waveform of the collector voltage. From [5].

5.3.2 Fitting I-V Measurements

First, we will study a method to fit the electrothermal collector current I-V model shown in (5.11) using measured dc values of V_{BE}, V_{CE}, and I_C.

$$\begin{aligned} i_c = {} & g_m v_{be} + K_{2GM} \cdot v_{be}^2 + K_{3GM} \cdot v_{be}^3 \\ & + g_o v_{ce} + K_{2GO} \cdot v_{ce}^2 + K_{3GO} \cdot v_{ce}^3 \\ & + K_{2GMGO} \cdot v_{be} \cdot v_{ce} + K_{3GM2GO} \cdot v_{be}^2 \cdot v_{ce} \\ & + K_{3GMGO2} \cdot v_{be} \cdot v_{ce}^2 \\ & + K_{2GMT} \cdot t_J + K_{3GMT} \cdot t_J \cdot v_{be} + K_{3GOT} \cdot t_J \cdot v_{ce} \end{aligned} \tag{5.11}$$

To solve the 12 electrothermal coefficients in (5.11) plus the dc operating point I_{DC} we need at least 13 measurements. To solve the coefficients we write the model functions (powers and cross-products of v_{be}

and v_{ce}) into matrix $\boldsymbol{M}$, the corresponding coefficients into vector $\boldsymbol{A}$, and the measured collector currents into vector $\boldsymbol{Y}$:

$$\boldsymbol{A} = [I_{DC}, g_m, K_{2GM}, K_{3GM}, g_o, K_{2GO}, K_{3GO}, K_{2GMGO}, K_{3GM2GO}, K_{3GMGO2}, K_{2GMT}, K_{3GMT}, K_{3GOT}]^T \tag{5.12}$$

(5.13)

$$\boldsymbol{M} = \begin{bmatrix} 1 & v_{i1} & v_{i1}^2 & v_{i1}^3 & v_{o1} & v_{o1}^2 & v_{o1}^3 & v_{i1}v_{o1} & v_{i1}^2v_{o1} & v_{i1}v_{o1}^2 & t_1 & t_1v_{i1} & t_1v_{o1} \\ 1 & v_{i2} & v_{i2}^2 & v_{i2}^3 & v_{o2} & v_{o2}^2 & v_{o2}^3 & v_{i2}v_{o2} & v_{i2}^2v_{o2} & v_{i2}v_{o2}^2 & t_2 & t_2v_{i2} & t_2v_{o2} \\ : & : & : & : & : & : & : & : & : & : & : & : & : \\ 1 & v_{iM} & v_{iM}^2 & v_{iM}^3 & v_{oM} & v_{oM}^2 & v_{oM}^3 & v_{iM}v_{oM} & v_{iM}^2v_{oM} & v_{iM}v_{oM}^2 & t_M & t_Mv_{iM} & t_Mv_{oM} \end{bmatrix},$$

and

$$\boldsymbol{Y} = \begin{bmatrix} I_1 & I_2 & .. & I_M \end{bmatrix}^T. \tag{5.14}$$

Now we can solve $\boldsymbol{A}$ from $\boldsymbol{MA}=\boldsymbol{Y}$ either exactly or using an LMSE fit. In $\boldsymbol{M}$, v_{ik} and v_{ok} are shorthand notations for the incremental voltages $v_{BE} - v_{BEQ}$ and $v_{CE} - v_{CEQ}$ of the kth measurement, around the chosen bias point v_{BEQ}, v_{CEQ}. The current vector is built to correspond to the desired junction temperature using linear interpolation, as explained in Section 5.2.2.

As pointed out in Section 5.1, the arrangement of the measurement points affects the numerical properties of this group of equations. Hence, we must find points that do not result in a linearly dependent group of equations, but still encompass the different nonlinearities as well as possible.

One almost orthogonal (but not minimal) way of choosing the locations of the measurement points is illustrated in Figure 5.7. First, the value of the transconductance and its nonlinearity can be obtained by setting the collector voltage to the quiescent point (0 VAC), measuring the collector current at four incremental v_{be} values, and fitting a third-degree polynomial of v_{be} to these points. With the same principle, the output conductance can be fitted by setting the base voltage to the quiescent point and by sweeping the collector voltage. Again, we have four measurement points, and nonlinearities up to the third degree can be fitted. Finally, cross-terms can be characterized using points C1-C4 that cover the corners of the v_{be}-v_{ce} fitting area. This reasoning calls for 12 measurement points to derive nine

electrical coefficients, and it can be used for sequential step-by-step extraction of the coefficients. However, the placement of the measurement points is quite well chosen to avoid numerical problems, and a simultaneous fit of all parameters can as well be done by placing these points into the ***M*** and ***Y*** matrix in (5.13) and (5.14) and by performing an LMSE fit.

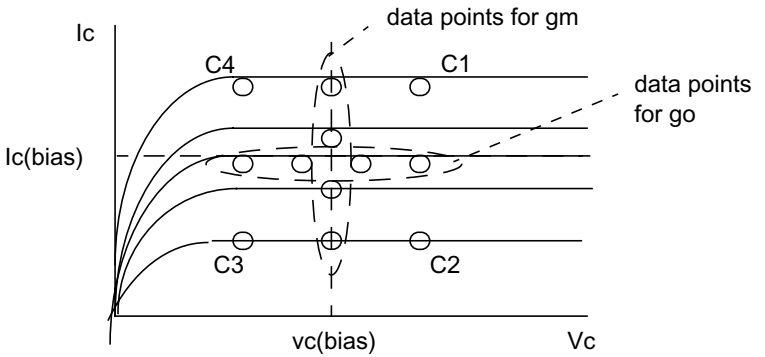


Figure 5.7 A dc characterization and the data points required for third-degree extraction. © IEEE 2001 [15].

5.4 AC Characterization Flow

In ac characterization, the conventional small-signal *S*-parameter measurements are performed using a network analyzer. The ac measurements are necessary for finding models for the capacitive nonlinearities, but they can be used to measure conductive nonlinearities, as well. The treatment of the measurement results is now much more complicated than in the dc measurements, however, as illustrated in Figure 5.8. It includes the calibration of the test fixture, which is normally implemented using reference impedances and the embedded software in the network analyzer. Then, isothermal (pulsed) *S*-parameter measurements are performed over a range of bias voltages and ambient temperatures. These results may still include the effects of the package that need to be de-embedded before extracting the small-signal circuit elements like g_m or C_{GS}. Finally, we need to know how to fit I-V and Q-V nonlinearities based on measured data on dQ/dv and dI/dv behavior.

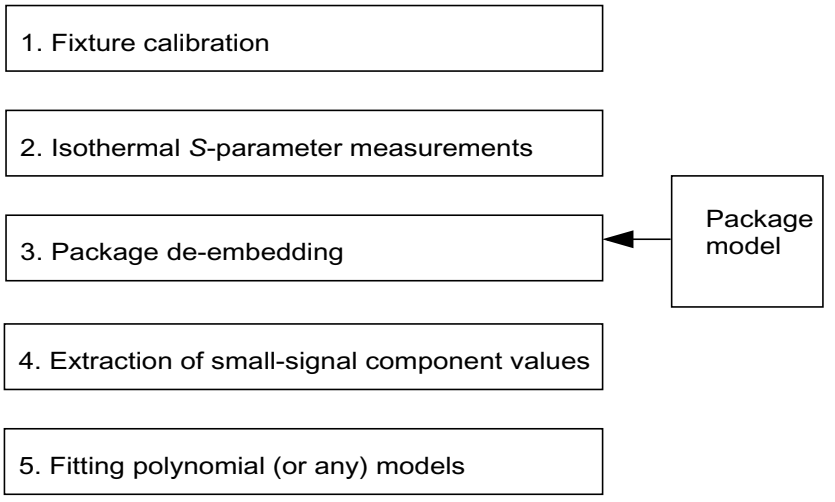


Figure 5.8 Characterization flow to obtain polynomial nonlinearity coefficients of the model by pulsed *S*-parameter measurements.

5.5 Pulsed *S*-Parameter Measurements

5.5.1 Test Setup

The pulsed *S*-parameters can be measured using the test setup shown in Figure 5.9. A network analyzer (NWA) is connected to the device under test (DUT) through dc-blockers to avoid the need of series capacitors on the test board. The drain bias voltage is fed through an inductance to make the bias impedance high enough at the RF frequency. Since the gate current is low, the gate bias is fed through a high-enough resistor to keep the bias impedance high. The measurement starts when the gate voltage is triggered to set the correct bias values. After that the NWA measures the *S*-parameters of the DUT at one frequency. The NWA used in the measurements reported here is capable of measuring all four *S*-parameters in 333 μs, and the results of three repeated measurements are averaged, resulting in a measurement time of 1 ms. At the same time, oscilloscope probes measure the corresponding bias values. As a result, all four *S*-

parameters at some value of drain and gate voltages and temperature are obtained in 1 ms, and by sweeping both the drain and gate voltages, pulsed *S*-parameters over the ranges of drain voltage and current are obtained. By repeating the measurements at two temperatures, the linear temperature dependencies of the *S*-parameters can also be obtained.

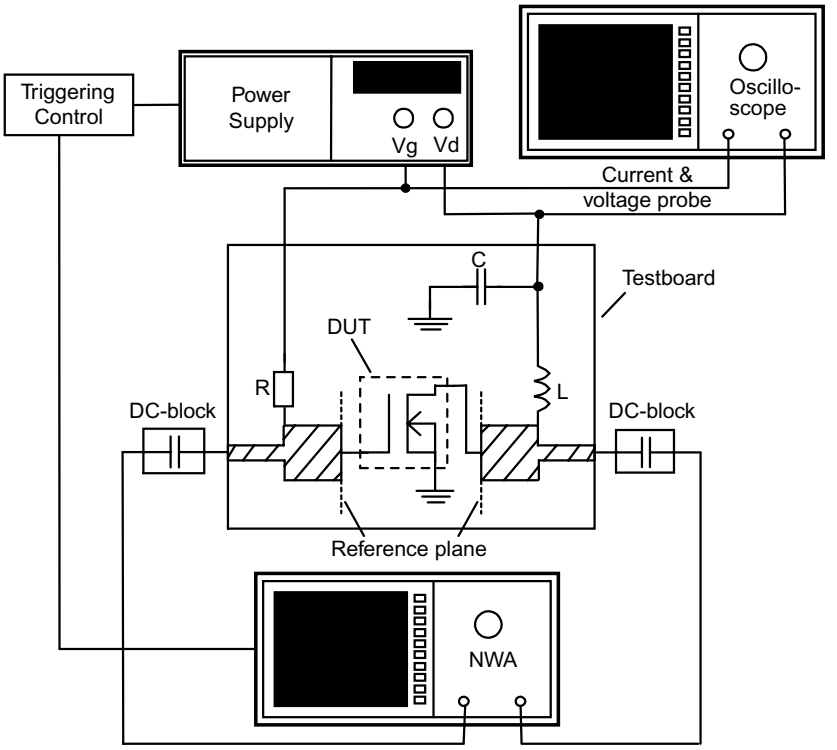


Figure 5.9 The test setup for pulsed *S*-parameter measurements. © IEEE 2002 [12].

When designing the test fixture, it is important to avoid electrical time-constants that are comparable to the pulse length. Large time constants slow down the electrical settling time, therefore potentially causing inaccuracies in the results. It is also important to measure the actual node voltages instead of voltages of the dc supplies. The series inductance especially may exhibit a noticeable resistance and consequently cause a significant voltage drop between supply and node voltages. Third, the duty cycle of the pulsing must be low enough: The average dissipated power is

simply $(t_{ON}/T)*V_D*I_D$, and to keep the self-heating below approximately 1%, the duty cycle must also be less than 1%. Fourth, the settling speed of the power supplies may not be sufficient for pulsing, and one may need to trigger the gate voltage using a series switch, instead. In this case, a pull-down resistor must be added at the gate line to guarantee that V_G does not remain floating when the series switch in the gate bias line is opened.

5.5.2 Calibration

Conductive and capacitive nonlinearities are usually recognized and separated from each other by studying the phase of the measured *S*- or *Y*-parameters. Hence, accurate calibration of the test fixture is crucial for the accuracy of the characterization. The measurements of packaged devices must often be made at the center frequency, and as an example, an error of 1 mm in a reference plane causes a phase error of 5º at 2.14 GHz on an $\varepsilon_r = 4$ substrate.

The other and even more severe problem is the impedance level inside the power transistor. If the output impedance of the transistor is 2 ohms, for example, the reflection coefficient is almost 1 using a 50-ohm reference, and the relative error for measuring these impedance levels might be as high as 10 to 20% [16]. When measuring high power devices, the use of on-board impedance transformers may be necessary [17].

The third problem arises from the nonidealities of the calibration standards. Usually the measurement is calibrated using short, reflect, load (50 ohm), through and isolation standards. At high frequencies, the 50-ohm load is not necessarily accurate, causing errors to calibration. This is the situation especially if homemade calibration boards are used instead of an accurate calibration kit. Since the test board presented in Section 6.2 includes dc bias feeds and other components, the use of home-made calibration boards is mandatory, introducing the problem caused by calibration standard inaccuracy. However, this problem can be mostly circumvented using the through-reflect-line calibration (TRL) [18, 19], and the homemade calibration standards similar to the original testboard can be used.

The TRL calibration is illustrated in Figure 5.10. The reflect is identical (except without the transistor) to the original testboard used in the measurements and the through is otherwise similar to the reflect but the gap between input and output reference planes is set to zero. The line is also similar to the through but the electrical distance between the DUT input and output is set to approximately 90º. All the calibration boards should include the same chip components as the original testboard, minimizing the calibration errors. However, due to statistical variation between chip

components, the use of them is not preferred on calibration boards; separate dc blockers can be used to avoid the use of onboard series capacitors.

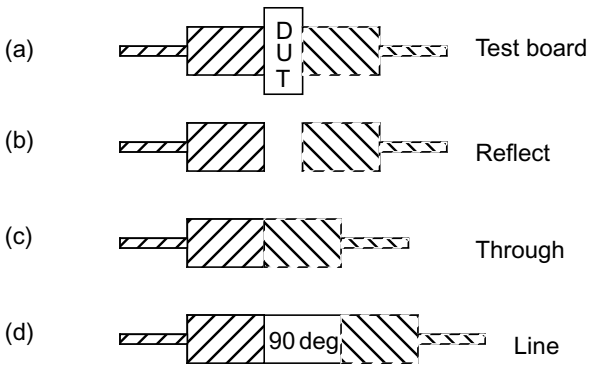


Figure 5.10 The test board (a) and calibration boards (b)-(d) needed for TRL-calibration.

5.6 De-embedding the Effects of the Package

The measured *S*-parameters of packaged transistors present not only the intrinsic transistor, but also the extrinsic part of it, consisting of lead inductances, lead resistances, and mutual coupling between the pins. The intrinsic transistor now has to be de-embedded from an extrinsic transistor before the model can be extracted. If the package can be modeled as plain series bond wire inductances and resistances, the values of these are quite easy to estimate from cold transistor measurements. However, the packages of high-power transistors are sometimes quite complicated, including on-chip (or in-package) matching circuitry, and their extraction is more complicated. For example, the in-package matching network of the MRF21030 30-W LDMOS is shown in Figure 5.11, including three chip capacitors and lots of bond wire inductances and mutual couplings. To obtain the parameters of the intrinsic transistor we present the de-embedding in the general and simplified cases.

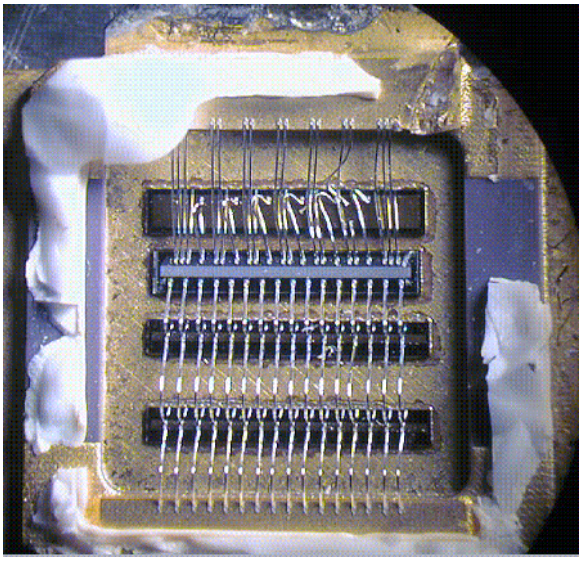


Figure 5.11 In-package matching network of the MRF21030 LDMOS.

5.6.1 Full 4-Port De-embedding

This section presents a de-embedding procedure that can be applied to any kind of package, no matter the type of the in-package matching circuitry. The technique was originally proposed for 16-term calibration of test fixtures in network analyzer measurements [20], but here the error box is used to model the transistor package instead. The only requirement – and major limitation – of this technique is that we need to have a full 4-port model of the package. In this example, it is available as part of the commercial simulation model.

The package model and the intrinsic grounded source FET are shown in Figure 5.12. A full 4-port, 16-term model with all mutual couplings is used to model the couplings between external and on-chip drain and gate terminals. Now the reflected waves b_i can be given as products of incident waves a_j and the *S*-parameters S_{ij} of the package. To simplify further notations, the 4x4 4-port *S*-parameter matrix is divided into four 2x2 sub-matrices $\boldsymbol{e_1}$ - $\boldsymbol{e_4}$, where $\boldsymbol{e_1}$ models connections between ports 1 and 4 (i.e., directly between external pins), $\boldsymbol{e_4}$ those between the on-chip drain and gate terminals, and $\boldsymbol{e_2}$ and $\boldsymbol{e_3}$ all the cross connections.

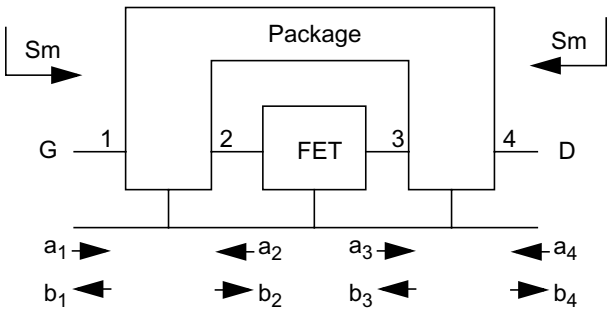


Figure 5.12 A 4-port model of the package. © IEEE 2002 [12].

$$\begin{bmatrix} b_1 \\ b_4 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} s_{11} & s_{14} & s_{12} & s_{13} \\ s_{41} & s_{44} & s_{42} & s_{43} \\ s_{21} & s_{24} & s_{22} & s_{23} \\ s_{31} & s_{34} & s_{32} & s_{33} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_4 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \boldsymbol{e_1} & \boldsymbol{e_2} \\ \boldsymbol{e_3} & \boldsymbol{e_4} \end{bmatrix} \cdot \begin{bmatrix} a_1 \\ a_4 \\ a_2 \\ a_3 \end{bmatrix} \tag{5.15}$$

Noting that ports 2 and 3 are, according to (5.16), interrelated by the *S*-parameters of the intrinsic transistor

$$\begin{bmatrix} a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} s_{11\text{int}} & s_{12\text{int}} \\ s_{21\text{int}} & s_{22\text{int}} \end{bmatrix} \cdot \begin{bmatrix} b_2 \\ b_3 \end{bmatrix} = \boldsymbol{S}_{\mathbf{int}} \cdot \begin{bmatrix} b_2 \\ b_3 \end{bmatrix} \tag{5.16}$$

we can (after some manipulation) solve the *S*-parameters of the intrinsic transistor simply by

$$\boldsymbol{S}_{\mathbf{int}} = (\boldsymbol{e_3} \cdot (\boldsymbol{S_m} - \boldsymbol{e_1})^{-1} \cdot \boldsymbol{e_2} + \boldsymbol{e_4})^{-1} \quad , \tag{5.17}$$

where $\boldsymbol{e_1}$ - $\boldsymbol{e_4}$ are the 2x2 submatrices given in (5.15) and $\boldsymbol{S}_{\mathbf{m}}$ are the measured 2-port *S*-parameter matrix of the packaged device.

The functionality of the de-embedding is demonstrated here by plotting the external and internal S_{11} parameter of the MRF21030 as a function of

frequency. Due to the in-package matching network, the extrinsic S_{11} plotted in Figure 5.13(a) is very frequency dependent, especially around the desired center frequency of 2.1 GHz. However, the intrinsic S_{11int} obtained using (5.17) is plotted in Figure 5.13(b), and it is reasonably wideband and free of resonances. This is expected, because S_{11int} is caused mostly by C_{GS} and C_{GD}, and the de-embedded result can now be used to extract values for C_{GS} and C_{GD}.

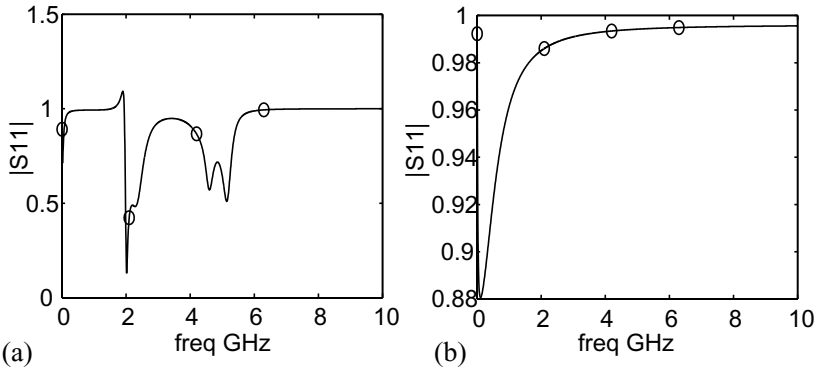


Figure 5.13 The simulated magnitude of S_{11} of (a) a packaged device, and (b) the intrinsic transistor, obtained after de-embedding.

5.6.2 De-embedding Plain Bonding Wires

If the package is simple and can be modeled by series inductors and resistors without mutual couplings, as shown in Figure 5.14(a), we can estimate the values of the series components from cold (pinched-off) transistor measurements. The idea of cold transistor measurements is that when pinched off, all conductive terms reduce to zero, and the transistor itself can be modeled by a purely capacitive pi or T network, shown in Figure 5.14(b). Since the intrinsic transistor reduces to simple capacitive network, the extraction of resistive and inductive parasitics becomes easy.

We start the procedure by removing the pin capacitances C_{GG} and C_{DD}. This is done simply by subtracting the values of $j\omega C_{GG}$ and $j\omega C_{DD}$ from the measured y_{11} and y_{22} parameters of the cold FET. After that, the lead resistances R_G, R_S, and R_D are the only resistive components, and they can be calculated directly from the real parts of the Z-parameters of the cold transistor [Z_C, surrounded by the dash-dot line in Figure 5.14(a)]:

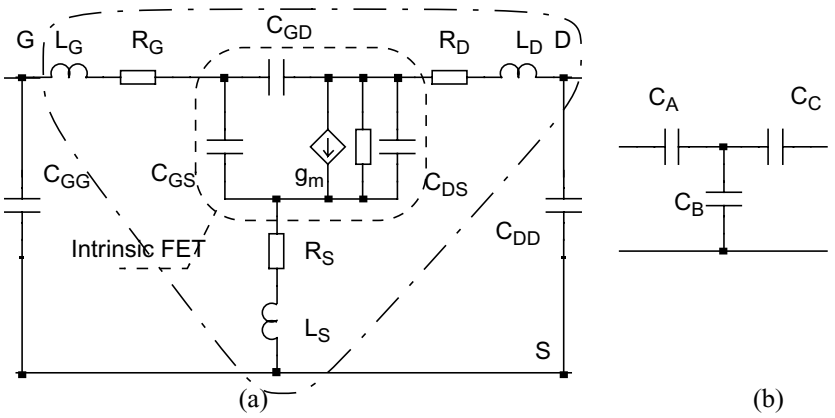


Figure 5.14 (a) An FET including extrinsic components, and (b) an intrinsic pinched-off cold FET. Modified from [21].

$$\begin{aligned} R_G &= Re(Z_{C11} - Z_{C12}) \\ R_S &= Re(Z_{C12}) = Re(Z_{C21}) \\ R_D &= Re(Z_{C22} - Z_{C12}) \end{aligned} \tag{5.18}$$

The lead inductances are a bit trickier to find, but their values can be found from the equations

$$\begin{aligned} \omega \cdot Im(Z_{C11}) &\approx \omega^2 \cdot (L_G + L_S) - \frac{1}{C_{AB}} \\ \omega \cdot Im(Z_{C12}) &\approx \omega^2 \cdot L_S - \frac{1}{C_B} \\ \omega \cdot Im(Z_{C22}) &\approx \omega^2 \cdot (L_D + L_S) - \frac{1}{C_{BC}} \end{aligned} \tag{5.19}$$

Here the values for C_{AB}, C_B, and C_{BC} need not be known, but the lead inductances can be calculated from the slope of the ω^2 - ω*Im(Z) curves [21].

Once we have extracted the package model from cold transistor measurements, we can now remove the package simply by subtracting it from the *Z*-parameters $\boldsymbol{Z}_{\mathbf{ext}}$ of the packaged device:

$$\boldsymbol{Z}_{int} = \boldsymbol{Z}_{ext} - \begin{bmatrix} Z_G + Z_S & Z_S \\ Z_S & Z_D + Z_S \end{bmatrix} \tag{5.20}$$

where Z_G, Z_S, and Z_G are the total series impedances at the gate, source, and drain, respectively.

5.7 Calculation of Small-Signal Parameters

The next step is to find the values of the equivalent small-signal model, based on intrinsic *S*-parameter values that are further converted to *Y*-parameters by solving the matrix equation (5.21), where $\boldsymbol{Y}$, $\boldsymbol{S}$, and $\boldsymbol{I}$ are the *Y*- and *S*-parameter and unity matrices, respectively, and Z_o is the reference impedance. A lot of different small-signal models exist, and most sophisticated intrinsic models include substrate coupling effects and transcapacitances, which are important above a few gigahertz, making the extracted nonlinear ac model valid up to 10 GHz [22, 23]. Since the example transistors considered in this book are used around 2 GHz, an approximate *Y*-parameter analysis is employed. As an example, a quasi-static FET pi model is presented in Figure 5.15.

$$\boldsymbol{Y} = \frac{1}{Z_o} \cdot (\boldsymbol{I} - \boldsymbol{S}) \cdot (\boldsymbol{I} + \boldsymbol{S})^{-1} \tag{5.21}$$

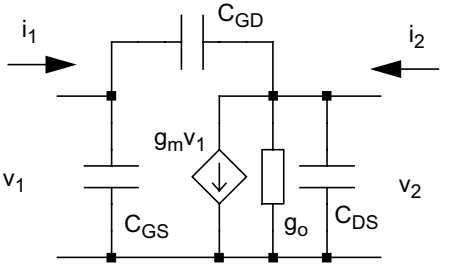


Figure 5.15 The FET pi model showing the components to be extracted.

The *Y*-parameters of this pi model can be found to be

$$\begin{bmatrix} i_1 \\ i_2 \end{bmatrix} = \begin{bmatrix} j\omega(C_{GS} + C_{GD}) & -j\omega C_{GD} \\ (g_m e^{-j\omega\tau} - j\omega C_{GD}) & (g_o + j\omega(C_{DS} + C_{GD})) \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \tag{5.22}$$

The circuit element values can be found by comparing (5.22) to the measured and de-embedded *Y*-parameter values. Actually there are many ways to solve the element values, some of them being less prone to errors in measurements than others, and the method suggested in [21] for calculating the capacitance values is used here. The feedback capacitance C_{GD} can be obtained by

$$C_{GD} = \frac{-Im(y_{12})}{\omega} \tag{5.23}$$

and the drain-to-source capacitance by

$$C_{DS} = \frac{Im(y_{22})}{\omega} - C_{GD}. \tag{5.24}$$

The value of C_{GS} could be calculated using (5.24) and replacing y_{22} with y_{11}. However, [21] suggests to calculate it as follows

$$C_{GS} = \frac{|y_{11} + y_{12}|^2}{\omega \cdot Im(y_{11} + y_{12})}. \tag{5.25}$$

while in a BJT also the real part caused by g_{pi} must be removed. The output conductance g_o can be written as

$$g_o = Re(y_{22}). \tag{5.26}$$

Basically, g_o could be measured at any frequency, but low-frequency measurements are usually the most accurate. However, a strongly reflective in-package matching or the dc blockers may force the measurement at the center frequency here, too.

Finally the transconductance could be written as the real part of y_{21}. However, this is not a very accurate way to calculate the g_m, because the propagation delay τ in g_m rotates the g_m term in y_{21}. A more accurate way to calculate the g_m is to subtract the effect of C_{GD} from the y_{21}, leaving just the g_m term. This can be formulated by

$$g_m = |y_{21} - y_{12}| . \tag{5.27}$$

Above a procedure for extracting the small-signal element values was presented. If data over a range of frequencies is available, the extraction should also be performed over a wide frequency range to check the correctness of the model, measurements, and de-embedding: the values of the extracted circuit elements should be independent of frequency. The real part of y_{11} and y_{12} can also be used as a figure of merit, because in a quasi-static operation of an FET, only the series resistances cause some amount of real part to these parameters.

5.8 Fitting the AC Measurements

5.8.1 Fitting of Nonlinear Capacitances

The required model for Q-V nonlinearities is a polynomial of charge. As the capacitance is easier to measure, we have to see how this affects our fitting procedure. As an example, the nonlinear C_{GS} is fitted. The desired model is of form

$$\begin{aligned} Q_{GS} = {} & C_{GS} \cdot v_{GS} + K_{2CGS} \cdot v_{GS}^2 + K_{3CGS} \cdot v_{GS}^3 \\ & + K_{2CGST} \cdot t_J + K_{3CGST} \cdot t_J \cdot v_{GS} \end{aligned} \tag{5.28}$$

and its derivative with respect to v_{GS} is the measurable capacitance C_{GS}

$$\frac{\partial Q_{GS}}{\partial v_{GS}} = C_{GS} + 2K_{2CGS} \cdot v_{GS} + 3K_{3CGS} \cdot v_{GS}^2 + K_{3CGST} \cdot t_J . \tag{5.29}$$

The relation between the capacitance and the charge is illustrated in Figure 5.16. As we need to fit three parameters (ignoring the temperature effects for a moment), we need at least three measurements at voltages V_{GS1}-V_{GS3}, marked as C_{GS1}-C_{GS3} in Figure 5.16(a). To include the thermal effects we need a fourth measurement point, at least one of which is measured at another temperature.

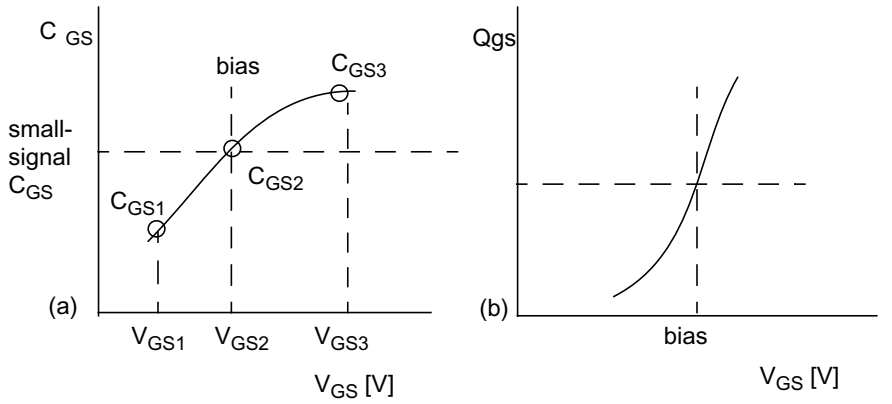


Figure 5.16 (a) C_{GS} and (b) Q_{GS} as a function of V_{GS}.

Based on (5.29), the measured C_{GS} values can be written as a group of equations (5.30), from which the vector of coefficients $[C_{GS}, K_{2CGS}, K_{3CGS}, K_{3CGST}]^T$ can be solved either exactly, or (if we have more measurement points) using an LMSE fit.

$$\begin{bmatrix} C_{GS1} \\ C_{GS2} \\ C_{GS3} \\ C_{GS4} \end{bmatrix} = \begin{bmatrix} 1 & 2v_{GS1} & 3v_{GS1}^2 & t_1 \\ 1 & 2v_{GS2} & 3v_{GS2}^2 & t_2 \\ 1 & 2v_{GS3} & 3v_{GS3}^2 & t_3 \\ 1 & 2v_{GS4} & 3v_{GS4} & t_4 \end{bmatrix} \cdot \begin{bmatrix} C_{GS} \\ K_{2CGS} \\ K_{3CGS} \\ K_{3CGST} \end{bmatrix} \tag{5.30}$$

Here v_{GSi} is again the difference between the measured V_{GS} voltage and the chosen bias point V_{GSQ} and $t_i = T_i - T_j$ is the incremental temperature of measurement i (at least one of them must be different from the others).

Other capacitances can be fitted in a similar manner. Note that the K_{2CGST} term (modeling the charge as a function of temperature) is lost in the differentiation and hence cannot be extracted using this method.

5.8.2 Fitting of Drain Current Nonlinearities

I-V nonlinearities are mostly characterized by using pulsed current measurements. It is, however, possible to use the technique described above to extract conductive nonlinearities as well from small-signal ac measurements. Similarly, we start from the current equation (5.11) and differentiate it with respect to v_{GS} and v_{DS} to obtain the measurable parameters g_m and g_o (thermal effects are ignored at the moment):

$$\begin{aligned}\partial i_D/\partial v_g &= g_m + 2K_{2GM}\cdot v_g + 3K_{3GM}\cdot v_g^2 \\ &+ K_{2GMGO}\cdot v_d + 2K_{3GM2GO}\cdot v_g\cdot v_d + K_{3GMGO2}\cdot v_d^2\end{aligned} \tag{5.31}$$

and

$$\begin{aligned}\partial i_D/\partial v_d &= g_o + 2K_{2GO}\cdot v_d + 3K_{3GO}\cdot v_d^2 \\ &+ K_{2GMGO}\cdot v_g + K_{3GM2GO}\cdot v_g^2 + 2K_{3GMGO2}\cdot v_g\cdot v_d\end{aligned} \quad . \tag{5.32}$$

These equations have nine unknown parameters, so that at least nine measurements are needed to solve all the coefficients. The arrangement of the measurement points is again quite free, but as an example the nine points can be nicely arranged into a grid shown in Figure 5.17, where point 5 is the chosen bias point and the rest are its eight neighboring values. At each point, the small-signal g_m and g_o values are measured.

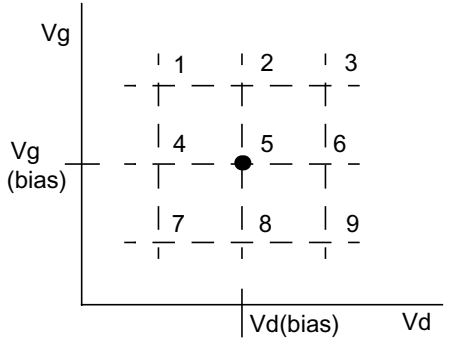


Figure 5.17 Selected bias point (5) and eight neighboring points for the calculation of drain current nonlinearities. © IEEE 2002 [11].

Thus, the matrix presentation of the measured data (ignoring the thermal terms for a moment) looks like $\boldsymbol{MC}=\boldsymbol{Y}$, where

$$\boldsymbol{M} = \begin{bmatrix} 1 & 2V_{\mathrm{G1}} & 3V_{\mathrm{G1}}^2 & 0 & 0 & 0 & V_{\mathrm{D1}} & 2V_{\mathrm{G1}}V_{\mathrm{D1}} & V_{\mathrm{D1}}^2 \\ 0 & 0 & 0 & 1 & 2V_{\mathrm{D1}} & 3V_{\mathrm{D1}}^2 & V_{G1} & V_{G1}^2 & 2V_{\mathrm{G1}}V_{\mathrm{D1}} \\ & & & & & \cdots & & & \\ 1 & 2V_{\mathrm{G9}} & 3V_{\mathrm{G9}}^2 & 0 & 0 & 0 & V_{\mathrm{D9}} & 2V_{\mathrm{G9}}V_{\mathrm{D9}} & V_{\mathrm{D9}}^2 \\ 0 & 0 & 0 & 1 & 2V_{\mathrm{D9}} & 3V_{\mathrm{D9}}^2 & V_{G9} & V_{\mathrm{G9}} & 2V_{\mathrm{G9}}V_{\mathrm{D9}} \end{bmatrix}, \quad (5.33)$$

$$\boldsymbol{Y} = \begin{bmatrix} g_{\mathrm{m1}} & g_{o1} & g_{\mathrm{m2}} & g_{o2} & \cdots & g_{\mathrm{m9}} & g_{o9} \end{bmatrix}^T, \quad (5.34)$$

and $\boldsymbol{C}$ is the vector of unknown coefficients:

$$\boldsymbol{C} = [g_{\mathrm{m}}, K_{2\mathrm{GM}}, K_{3\mathrm{GM}}, g_o, K_{2\mathrm{GO}}, K_{3\mathrm{GO}}, K_{2\mathrm{GMGO}}, K_{3\mathrm{GM2GO}}, K_{3\mathrm{GMGO2}}]^T \quad (5.35)$$

Since both g_{m} and g_{o} are measured in all nine data points we have altogether 18 equations and 9 unknown coefficients, which allow for LMSE solution of $\boldsymbol{C}$:

$$\boldsymbol{C} = (\boldsymbol{M}^T \cdot \boldsymbol{M})^{-1} \cdot (\boldsymbol{M}^T \cdot \boldsymbol{Y}) \quad (5.36)$$

The excess data in $\boldsymbol{M}$ can be exploited in several ways. First, it provides some filtering, and we can even remove some points that give the worst fit to the polynomial. Then, we can use the data to fit the electrothermal terms as well, provided that some of the points are measured at another temperature. Again, $K_{2\mathrm{GMT}}$ corresponding to a plain t_{J} term cannot be seen in the g_{m},g_{o} data, but $K_{3\mathrm{GMT}}$ and $K_{3\mathrm{GOT}}$ can be fitted by adding terms t_{J} and $v_{\mathrm{D}}{}^{*}t_{\mathrm{J}}$ to $\mathrm{g_m}$ data and $v_{\mathrm{G}}{}^{*}t_{\mathrm{J}}$ and t_{J} to g_{o} models in the $\boldsymbol{M}$ matrix. Alternatively, these coefficients can be extracted as a temperature dependence of the fitted $\mathrm{g_m}$ and $\mathrm{g_o}$ coefficients, for example, as

$$K_{3\mathrm{GOT}} = \frac{\overline{g_o(T_2)} - \overline{g_o(T_1)}}{(T_2 - T_1)}, \tag{5.37}$$

where T_1 and T_2 are the measurement temperatures and the g_o values are average values over the range of drain voltage extraction. To extract $K_{2\mathrm{GMT}}$, current measurements are necessarily needed.

It is worth noting that similar to Figure 5.16, where the Q-V curve was reconstructed by integrating the measured C-V curve, it is possible to reconstruct the I-V curve from the measured g_m, g_o data, provided that we have a dense enough grid of measurement points. This is illustrated in Figure 5.18. The figure also illustrates one major difference between the dc and ac characterization. In the dc measurements, all nonlinearity coefficients starting from the linear terms g_m and g_o must be derived from the I-V data (performing a numerical differentiation, in principle). In the ac measurements, g_m and g_o are already *measured* quantities, and the order of the fitted model is hence lower by one. Hence, less data points are needed, and, presumably, the fitting is less sensitive to numerical errors. However, the calibration, de-embedding, and so forth contribute their share into the measurement errors.

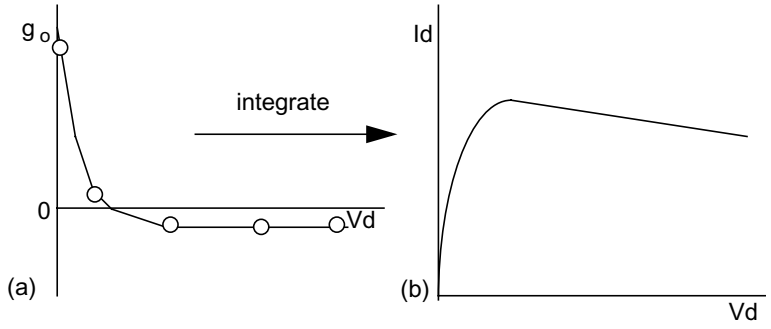


Figure 5.18 (a) Extracted g_o values, and (b) the corresponding drain current as a function of drain voltage. © IEEE 2001 [15].

5.9 Nonlinear Model of a 1-W BJT

In this section, the dc I-V characterization is applied to find the nonlinearity coefficients of the conductive nonlinearities in Phillips BFG11 power BJT. The nonlinearities of the capacitances (used in Chapter 4) are calculated from the model equations and SPICE parameters available. An external base resistor shown in Figure 5.6 is used in the measurement to reduce sensitivity of the input control.

The measured base and collector currents of BFG11 BJT are shown as functions of the input and collector voltages in Figure 5.19. The base current is independent of the collector voltage at low current levels, although a slight dependence is observed in the high current – high voltage region. This may be caused by self-heating, as the measurement time is as long as 10 ms. The base current is still quite independent of the collector voltage, but the same conclusion cannot be reached concerning the collector current. Estimated from Figure 5.19(b), the Early voltage of the device is as low as 8V to 12V, and strong cross-terms are needed for modeling the output behavior.

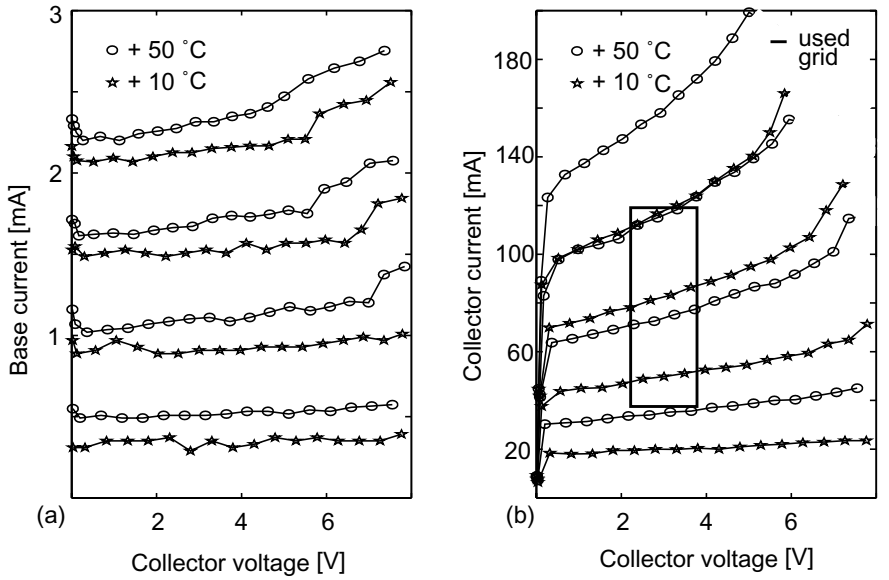


Figure 5.19 Measured (a) base current and (b) collector current of a BFG11 at two temperatures. The fitting range is enclosed by the box. From [5].

The nonlinearity coefficients of BFG11 BJT are extracted at collector and base bias voltages of 3V and 734 mV and a collector voltage and current swing of 1.5 Vpp and 80 mApp. The extraction range is shown in Figure 5.19(b).

The base current as a function of base voltage is plotted in Figure 5.20(a), showing an almost exponential electrical relationship between the two. As the vertical distance of the I_B curves in Figure 5.19(a) remains practically independent of the value of V_{BE}, the effect of K_{2GPIT} seems to dominate over K_{3GPIT}.

Since the collector current is measured as a function of the input and not of the base voltage, it also includes the nonlinearity of the g_{pi}, which has to be taken into account when extracting the nonlinear transconductance. The effects of g_{pi} can easily be distinguished by

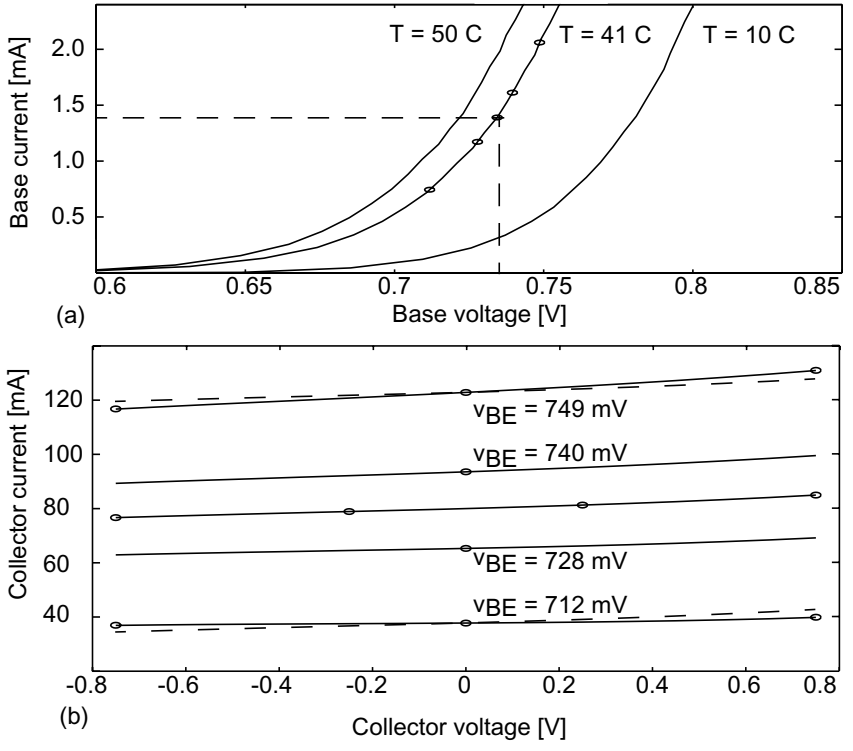


Figure 5.20 (a) Base current as a function of base voltage, and (b) collector current as a function of base and collector voltages. From [5].

mapping the measured data from the input voltage to the base voltage, after which the first three nonlinearity coefficients of (5.11) can be extracted at zero collector voltage (V_{CEQ}), causing all the other electrical terms to g_o to zero. Similarly, the nonlinear output conductance can be extracted at a zero base voltage (V_{BEQ}) because all the terms related to v_{BE} are zero. The cross-terms describe how the curves for the different base voltage differ in shape. Four measurement points, one in each corner of the I-V plane are used to characterize the three cross-terms, and the dashed lines in Figure 5.20(b) represent the situation where the cross-terms are zero. This causes up to 10% error in the collector current, clearly illustrating that the cross-terms must be included in the Volterra model.

The electrothermal terms related to the collector current can be extracted as follows: the second-degree term K_{2GMT} describes the current offset caused by the temperature at the bias point, while the two third-degree terms K_{3GMT} and K_{3GOT} are either functions of v_{BE} or v_{CE}, causing dependence of the current on changes in v_{BE} or v_{CE} and temperature.

The extracted nonlinearity coefficients are summarized in Table 5.1. The normalized nonlinearities of the transconductance and base conductance are close to each other, which implies that the nonlinearities are similar in shape and that these two nonlinearity mechanisms may partially cancel each other. In other words, even if the voltage signal at the base is highly distorted, the collector current may be fairly linear, and consequently the bipolar amplifiers can achieve high linearity even though the nonlinearities inside the device are exponential.

Note that the higher degree coefficients of each nonlinearity are normalized with the value of the linear term, and the cross-terms are normalized by g_o.

Table 5.1

Nonlinearity Coefficients for BFG11 at Vbase=734 mV and Vcoll=3V

	1st-degr.	2nd/1st	3rd/1st	T2/1st	T3/1st
g_m	2.4	14	160	0.00027	0.0051
g_{pi}	0.038	13	120	0.0016	–
g_o	0.0047	0.31	0.33	–	0.0098
Cross-terms	K_{2GMGO} = 0.23	K_{3GM2GO} = 3.2	K_{3GMGO2} = 0.017		

5.10 Nonlinear Model of a 1-W MESFET

In this section, the *S*-parameter characterization method is employed to determine both the capacitive and conductive nonlinearities of an Infineon CLY2 GaAs MESFET [24]. The *S*-parameters are measured in pulsed form over a range of bias conditions at temperatures of 0º and 50ºC, and the series bondwire impedances have been de-embedded before extracting the small-signal elements. Now the nonlinearity coefficients, shown in the following four figures, are extracted as functions of the selected bias point to illustrate how the nonlinearity of the circuit elements depend on the selection of the bias point. In all these figures, the *x*-axis represents V_{DSQ} in volts and *y*-axis I_{DSQ} in milliamperes, and the corresponding IM3 contours are shown in Chapter 4 in Figure 4.22. The IM3 vector plot of the amplifier based on this transistor model was analyzed in Chapter 4 using the bias point V_{DSQ}=4V and I_{DSQ}=150 mA.

The first column in Figure 5.21 corresponds to the first-, second-, and third-degree nonlinearities in the transconductance [the first row in (5.11)], while the second column shows the nonlinearity in the output conductance. Again, the higher degree coefficients are normalized with the linear terms. The shape of the nonlinearity of the transconductance is relatively independent of the drain voltage, and the normalized nonlinearity in transconductance is in the range 2% to 20% and decreases with increasing bias current. Similarly, the shape of the output conductance is relatively independent of the drain current, except in the ohmic region at low V_{DSQ} values, where the g_o varies rapidly.

Columns 1 and 2 in Figure 5.22 give the cross-terms and electrothermal terms. The increasing distortion below 2.5 V V_{DSQ} voltages in Figure 4.22 strongly correlate with the increasing g_o and cross-term nonlinearity. The electrothermal K_{2GMT} term is about 0.1 mA/K, meaning that a 1º variation in the junction temperature will cause a 0.5-mV or –60-dBc IM2 tone in the 5-ohm (at 2 MHz) drain bias impedance, but it will further attenuate when mixing up in K_{2GO}, for example. The effect of the electrothermal gain fluctuation K_{3GMT} is of the same order of magnitude.

The nonlinear capacitances are collected in Figure 5.23, where the left column shows the behavior of C_{GD}. The nonlinearity of the reverse biased Schottky junction is quite weak, but it must be remembered that the nonlinear current of C_{GD} is injected to the gate and is amplified to a significant degree in most cases, as a result of which even weak nonlinearity will cause quite a lot of distortion. For example, the amount of nonlinearity in C_{DS}, given in Figure 5.24, is similar to that in C_{GD} and the voltages across the two nonlinear capacitors are similar to each other, but

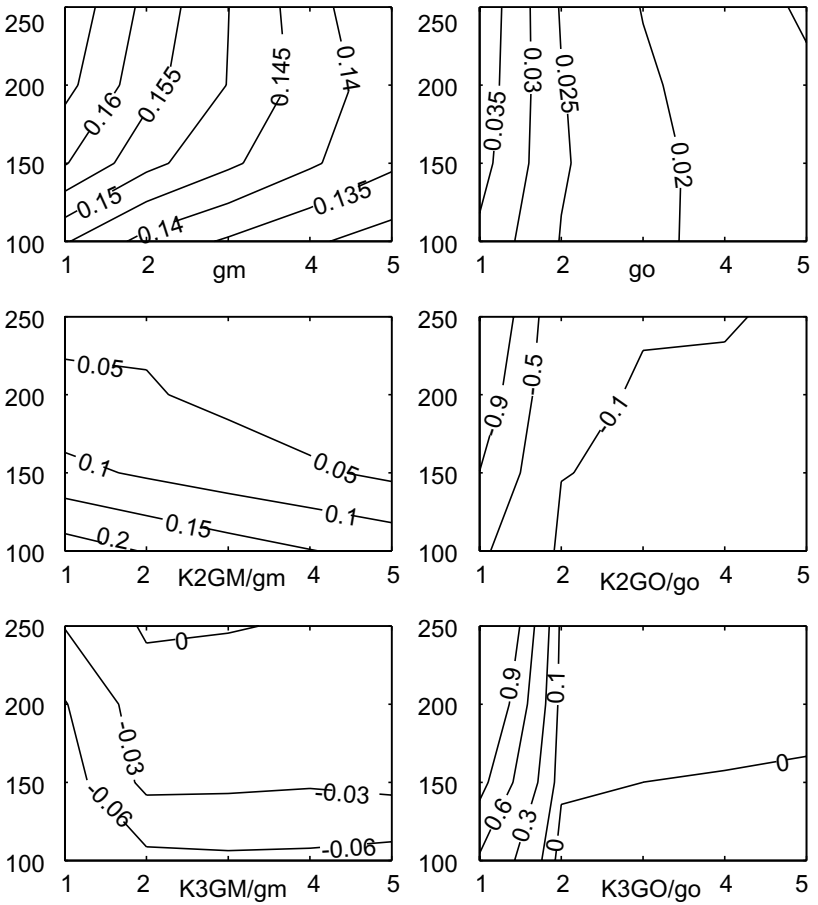


Figure 5.21 Transconductance and output conductance (vertical Id [mA], horizontal Vd [V]). The left column gives coefficients to terms g_m*v_{GS} + K_{2GM}*v_{GS}^2 + K_{3GM}*v_{GS}^3 and the right column to terms g_o*v_{DS} + K_{2GO}*v_{DS}^2 + K_{3GO}*v_{DS}^3. The displayed values of higher degree coefficients are normalized by the local values of g_m and g_o. © IEEE 2001 [15].

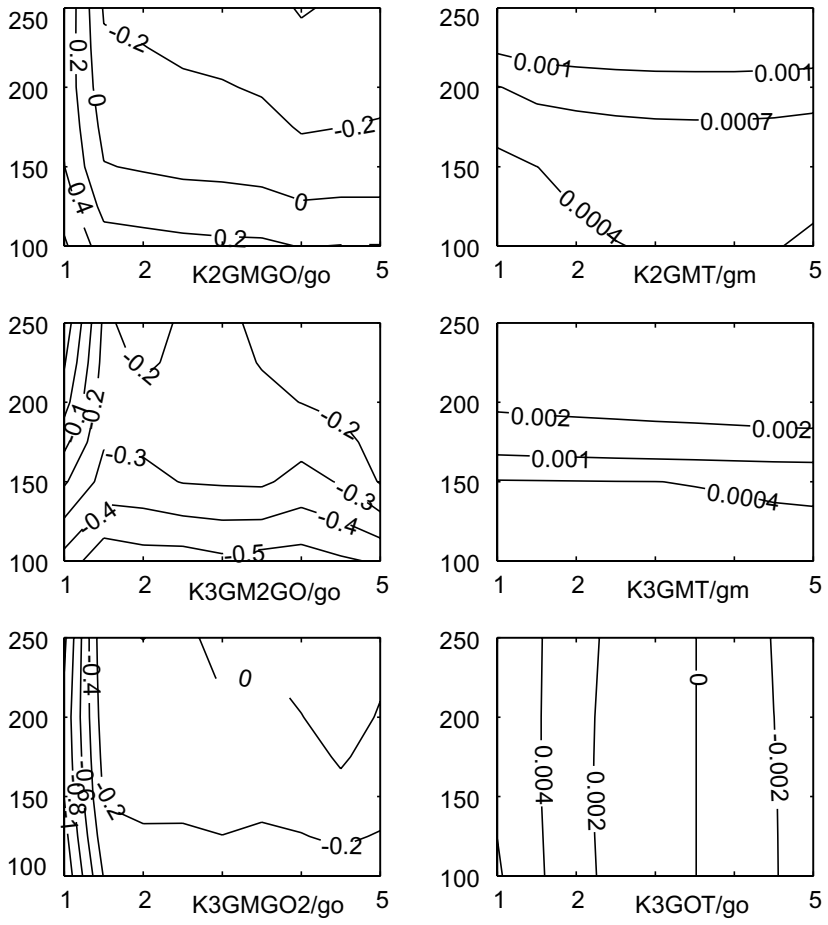


Figure 5.22 Cross-terms and electrothermal terms. The left column gives coefficients to terms K_{2GMGO}*v_{GS}*v_{DS} + K_{3GM2GO}*v_{GS}^2*v_{DS} + K_{3GMGO2}*v_{GS}*v_{DS}^2 and the right column to terms K_{2GMT}*t + K_{3GMT}*v_{GS}*t + K_{3GOT}*v_{DS}*t. © IEEE 2001 [15].

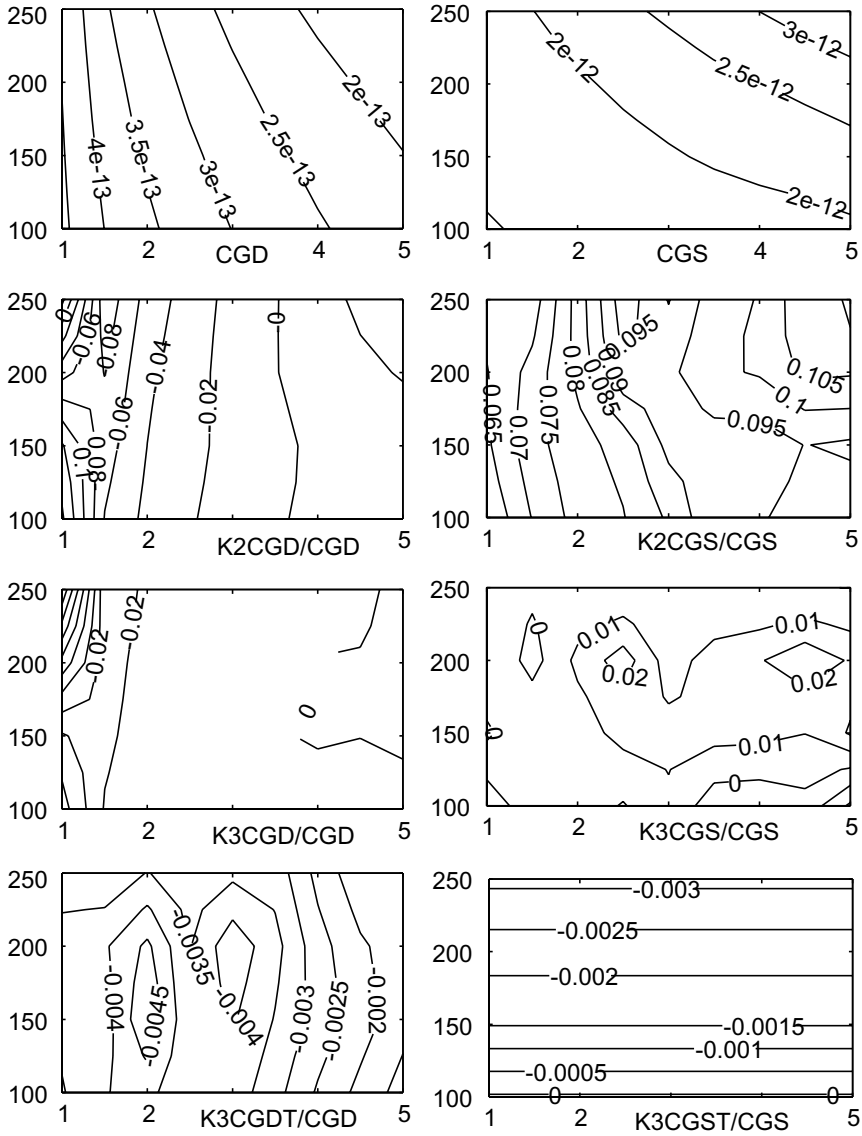


Figure 5.23 Electrical and electrothermal coefficients of C_{GD} (left) and C_{GS} (right). © IEEE 2001 [15].

the nonlinearity of C_{DS} affects only the output of the amplifier. For this reason the nonlinearity of C_{GD} can be more serious than that of C_{DS}.

The right column in Figure 5.23 presents the model for C_{GS}. The amplifying effect of input node distortion can again be seen. The normalized K_{3CGS} is smaller than K_{3GM} at V_{DS}=4V, I_{DS}=150 mA, and the amplitude of the IM3 current generated by it is much smaller than the IM3 current generated in K_{3GM}. However, the current generated in C_{GS} is injected to a higher total impedance at the gate and further amplified in g_m, and finally the IM3 contribution of K_{3GS} is actually slightly higher than the one generated in K_{3GM}, as seen from Figure 4.23 in Chapter 4.

For reference, the nonlinearity coefficients at the bias point V_{DS} = 4V, I_{DS} = 150 mA are collected in Table 5.2. The values of the higher degree coefficients are normalized by the linear term.

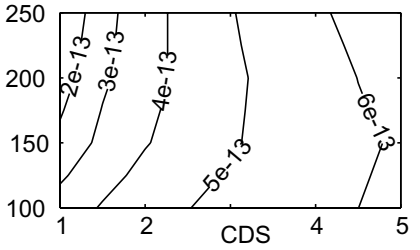


Figure 5.24 The value of C_{DS} versus bias point. © IEEE 2001 [15].

Table 5.2
Nonlinearity Coefficients for CLY2 at V_{DS} = 4V and I_D = 150 mA

	1st-degr.	2nd/1st	3rd/1st	T2/1st	T3/1st
g_m	0.14 S	0.07	-0.03	0.0006	0.0004
g_o	0.018 S	-0.07	0.0	–	-0.001
Cross-terms /g_o	K_{2GMGO} = -0.05	K_{3GM2GO} =-0.3	K_{3GMGO2} = 0.017		
C_{GD}	2.3 pF	-0.009	0		-0.003
C_{GS}	2.1 pF	0.095	-0.01		-0.0015

5.11 Nonlinear Model of a 30-W LDMOS

This section presents the extracted nonlinearity coefficients of the Motorola MRF 21030 power LDMOS transistor. In principle, the same *S*-parameter based characterization flow is followed as in the MESFET case in Section 5.10, but here the characterization is done completely using a circuit simulator, taking the measurement data from a MET device model provided by Motorola. Most simulators allow one to plot the small signal component values directly, but to test the 4-port de-embedding at the same time, the extraction starts here, too, from the *S*-parameter measurements of a packaged and biased device. Calibration procedures are naturally not needed in the simulator environment, and the debugging of the characterization routines is much easier when the results are noise-free and tractable. A simulation-based Volterra modeling is a very quick way of starting the Volterra analysis, provided that we have device models that we can trust; the extracted parameters are at most as accurate as the simulation model.

One thing that needs consideration is how to obtain isothermal measurements, as normally the MET model includes self-heating effects, and increasing bias would increase the junction temperature. Some simulators allow ac analysis on top of transient analysis, in which case pulsed measurements can be imitated. An easier approach is to set the thermal resistance of the model to zero and use the ambient temperature to force the junction temperature. In this way, the simulations are performed at the drain and gate voltage values of 2 V to 40 V and 3.0 V to 4.5V at the temperatures of 0º and 75ºC.

Next the steps of de-embedding, calculation of small-signal elements and fitting of nonlinearity coefficients are done as described in Sections 5.6 to 5.8. For example, the package de-embedding is illustrated in Figure 5.13.

The nonlinearity coefficients of MRF21030 as functions of the bias point are presented in the following three figures, where the *x*-axis represents V_{DS} and *y*-axis the V_{GS} bias voltage. In Figure 5.25, the first column corresponds to the linear, quadratic, and cubic nonlinearity of the transconductance. We can see that the K_{3GM} can be made zero at V_{GS} bias of 3.95 V. However, the g_m still has quite a strong square-law nonlinearity. The second column shows the nonlinearity in the output conductance that appears to be quite linear.

The left and right columns in Figure 5.26 give the cross-terms and electrothermal terms. The cross-terms are again quite large and make a large contribution to the total distortion. From the electrical terms K_{2GMT} looks insignificant (some microamperes/kelvin), but according to K_{3GMT} a

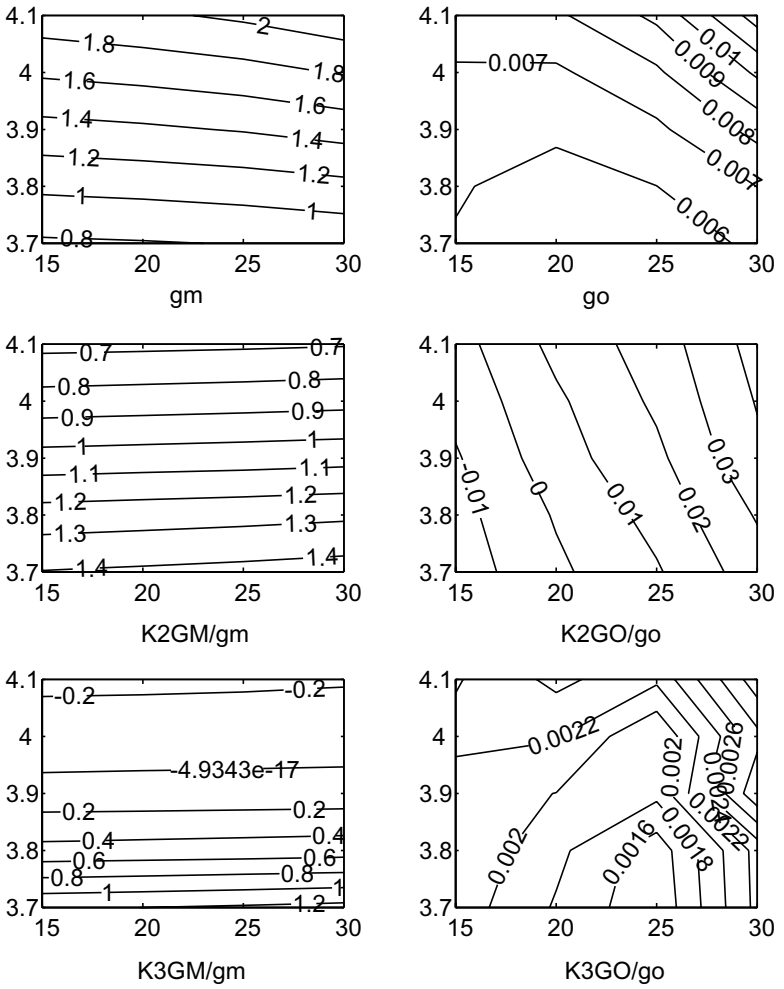


Figure 5.25 Transconductance and output conductance (vertical V_{GS} [V], horizontal V_{DS} [V]). The left column gives coefficients to terms $g_m{*}v_{GS}$ + $K_{2GM}{*}v_{GS}^2$ + $K_{3GM}{*}v_{GS}^3$ and the right column to terms $g_o{*}v_{DS}$ + $K_{2GO}{*}v_{DS}^2$ + $K_{3GO}{*}v_{DS}^3$. The displayed values of higher degree coefficients are normalized by the local values of g_m and g_o.

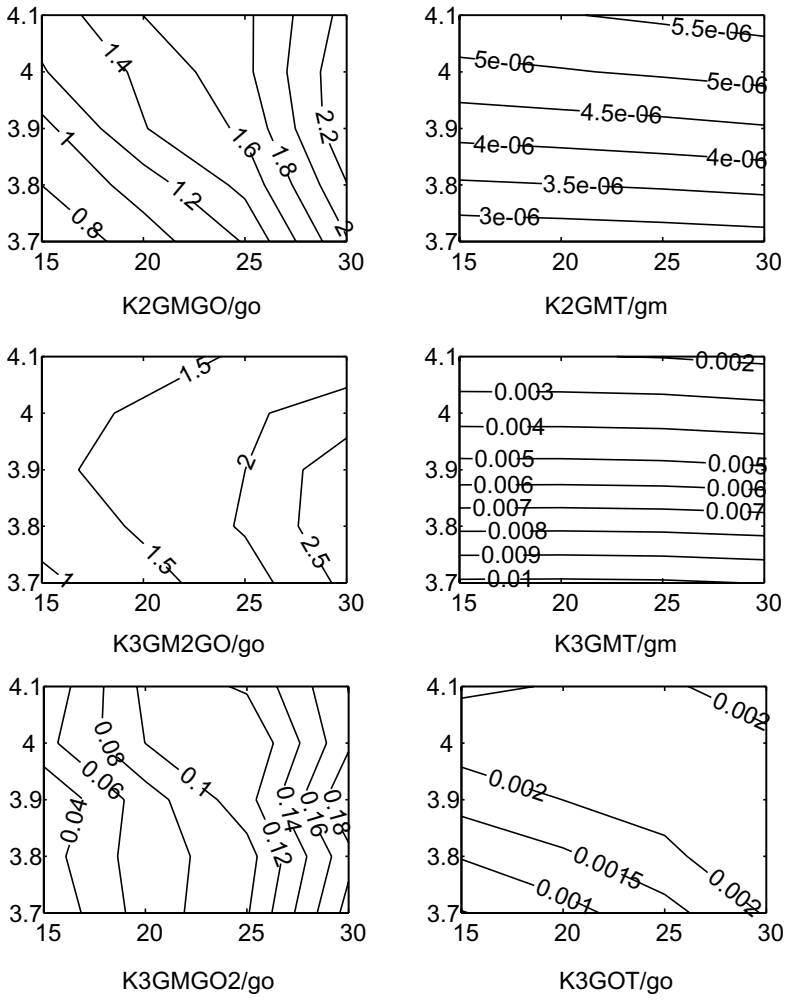


Figure 5.26 Cross-terms and electrothermal terms of MRF21030. The left column gives coefficients to terms $K_{2GMGO}*v_{GS}*v_{DS} + K_{3GM2GO}*v_{GS}^2*v_{DS} + K_{3GMGO2}*v_{GS}*v_{DS}^2$ and the right column to terms $K_{2GMT}*t + K_{3GMT}*v_{GS}*t + K_{3GOT}*v_{DS}*t$.

temperature fluctuation of 1ºC causes 0.5% to 1% modulation in the value of g_m, which will be seen in the IM3 sidebands.

The capacitances are shown in Figure 5.27. C_{GS} is also modeled as a nonlinear, two-dimensional function of the gate-to-source voltage and temperature. C_{GS} has moderate square-law and cubic nonlinearities. Again, K_{3GS} directly generates IM3 current to the gate, and it will be amplified by g_m*Zgate(@fundamental) to the output. The effect of K_{2CGS} can be affected by the baseband and second harmonic gate node impedances, and it generates small (due to jω dependency) envelope current and much larger second harmonic current.

In addition, C_{DS} and C_{DG} are also slightly nonlinear, but since their effects to the distortion are small, these nonlinearities are not presented here. However, their absolute values are shown in Figure 5.27.

As an example, the polynomial coefficients at the bias point V_{DS} = 28V, V_{GS} = 4V are listed in Table 5.3. For example, the g_m of the LDMOS has rather strong quadratic nonlinearity. The cross-terms also look strong, but as they are normalized by the relatively small g_o, they are still small compared to the nonlinearity of the g_m.

Table 5.3

Nonlinearity Coefficients for MRF21030 at V_{DS} = 28V and V_{GS} = 4V

	K_1	K_2/K_1	K_3/K_1	K_{2T}/K_1	K_{3T}/K_1
g_m	1.8 S	0.85	-0.1	5e-6	0.0035
g_o	0.04 S	0.035	0.002	–	0.002
Cross-terms /go	K_{2GMGO} = 2	K_{3GM2GO} =2	K_{3GMGO2} = 0.16		
C_{GS}	50 pF	0.08	-0.01		3e-5

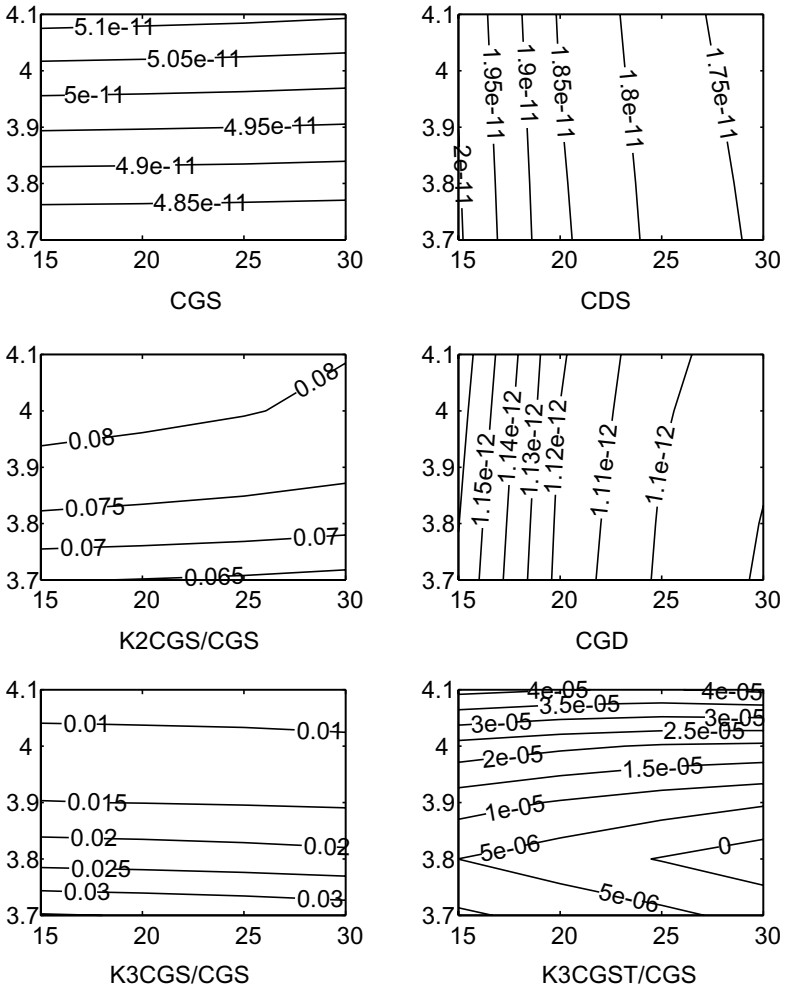


Figure 5.27 Nonlinearity of C_{GS}, and absolute values of C_{DS} and C_{GD} (vertical V_{GS} [V], horizontal V_{DS} [V]).

5.12 Summary

In this chapter, different characterization techniques for building Volterra models were presented, and as examples, the fitted nonlinearities of BJT, MESFET, and LDMOS transistors were shown. The I-V nonlinearity of the BJT was fitted using measured I-V data, while both FET type transistors were fitted using ac measurements, based on measured or simulated small-signal data.

Two general topics were discussed in the text. The first one was the selection of the fitting area, as it affects the accuracy of the modeling, numerical sensitivity of the fitting, and also the method of fitting. Both exact fitting with minimum amount of data and LMSE fitting with a larger amount of data was used. To estimate the required fitting range, the impedance levels and desired power levels must be known.

Self-heating is another important factor in nonlinear characterization of high-power devices, because nonlinearities arising from changes in terminal voltages and temperature are very difficult to separate from each other in steady-state measurements. Pulsed measurements with a low duty cycle are therefore used to investigate the transistor under as constant temperature conditions as possible. The effects of optimum pulse length were discussed, and it was noted that the pulse must be wide enough to produce an electrical steady-state, while at the same time it must be as short as possible to avoid self-heating.

The dc I-V measurements are quite straightforward, and the only technical problem is related to achieving isothermal measurements, which is normally done by pulsing the device on with a low duty cycle. The dc current measurements are always needed for obtaining all the electrothermal terms. However, capacitive nonlinearities must always be characterized using ac measurements.

The ac characterization flow starts from calibration issues. TRL calibration is used to avoid the need of accurate 50-ohm references, for example, as three slightly modified test boards are used as the calibration standards. Accurate calibration is very important, because low impedance, high-frequency measurements are prone to errors in calibration.

Since the measurements are performed for packaged transistors, the intrinsic part of it has to be calculated for polynomial extraction. This procedure is called de-embedding and it requires that either the model of the package already exists, or the extrinsic part of the transistor is simple enough so that each terminal can be considered as a series connection of R and L. Whenever neither of them is possible, the approaches presented here cannot be used. When the package is known or can be calculated, the

intrinsic part can be obtained using a 16-term error model, which also takes the cross couplings between the input and output into account.

Once the *S*-parameters of the intrinsic part of the transistor are obtained, the small-signal circuit elements can be calculated. This is done by comparing the measured and derived *Y*-parameters of the small-signal model to each other, as a result of which the equations for small-signal circuit elements can be deduced. The measurements of *S*-parameters and de-embedding are performed over a range of bias voltages and temperature, as a result of which the small-signal elements at different operating points are obtained.

The nonlinearities of each individual circuit element are now calculated based on changes in small-signal elements, and also the conductive nonlinearities can be fitted based on the measured g_m and g_o data. The greatest advantage of the ac method is that first derivatives of I-V curves are *measured*, not calculated quantities, which means that less data is needed in the fitting process. As an example, let us consider that the measured dc values are 100 mA and 105 mA at the V_D of 3V and 4V. The linear output conductance can then be calculated to be 5 mS. Let us further consider the accuracy of the current measurements to be 1%, which gives worst-case values for g_o to be 3 and 7 mS, which corresponds the errors up to 40%. Thus, the dc method causes errors even to the linear element values if small errors in data points exist, and this error is amplified when extracting higher degree coefficients [25]. The dominant error sources in ac analysis, on the other hand, are the accuracy of calibration, and de-embedding.

The Volterra model can also be characterized by means of a circuit simulator, in which the device model can be simulated in a manner similar to the measurements presented in this chapter, enabling a full Volterra model to be characterized. This is a very easy approach compared with measurements, but the problem is that the extracted nonlinearity coefficients cannot be more accurate than the derivatives of the model equations.

5.13 Key Points to Remember

1. The polynomial model can be fitted locally to the existing data, to an area set by the bias point and the estimated signal swing. The fitting range and placement of the measurement points affect the accuracy of the fitting.

2. To avoid self-heating, pulsed measurements with a low duty cycle and short enough pulses must be used.
3. Capacitive components must be measured using pulsed *S*-parameter measurements. Conductive components can be characterized either with pulsed dc measurements or pulsed ac measurements.
4. Once the parameters of the packaged device are measured, the effects of the package must be removed by the procedure called de-embedding. The de-embedding presented here requires that the model of the package exists, or it can be modeled by plain series RL networks.
5. Based on measured and de-embedded *S*-parameters, the small-signal element values of the model over the range of bias values can be calculated.
6. Nonlinearities of the circuit elements can be calculated based on fitting polynomial functions to the small-signal element data, presented as a function of terminal voltages and junction temperature.

References

[1] Maas, S., and A. Crosmun, "Modeling the gate I/V characteristic of a GaAs MESFET for Volterra-series analysis," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 37, No. 7, 1989, pp.1134-1136.

[2] Pedro, J., and J. Perez, "Accurate simulation of GaAs MESFET's intermodulation distortion using a new drain-source current model," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 42, No. 1, 1994, pp. 25-33.

[3] Sobhy, M., et al., "Nonlinear system and subsystem modeling in time domain," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 44, No. 12, 1996, pp. 2571-2579.

[4] Clark, C., et al., "Time-domain envelope measurement technique with application to wideband power amplifier modeling," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 46, No. 12, 1998, pp. 2531-2540.

[5] Vuolevi, J., "Analysis, measurement and cancellation of the bandwidth and amplitude dependence of intermodulation distortion in RF power amplifiers," Doctoral thesis, University of Oulu, Oulu, Finland, 2001.

[6] Veijola, T., M. Andesson, and A. Kallio, "Parameter extraction procedure for an electrothermal transistor model," *Proc. BEC'96*, Tallinn, Estonia, pp. 71-72.

[7] Veijola, T., and M. Andesson, "Combined electrical and thermal parameter extraction for transistor model," *1997 European Conference on Circuit Theory and Design,* Budapest, Hungary, pp. 754-759.

[8] Parker, A., et al., "Determining timing for isothermal pulsed-bias S-parameter measurements," *IEEE 1996 MTT-S International Microwave Symposium Digest*, Vol. 3, pp. 1707-1710.

[9] Collantes, J., et al., "A new large-signal model based on pulse measurement techniques for RF power MOSFET," *IEEE 1995 MTT-S International Microwave Symposium Digest*, Vol. 3, pp. 1553-1556.

[10] Batty, W., et al., "Fully physical time-dependent compact thermal modelling of complex non linear 3-dimensional systems for device and circuit level electro-thermal CAD," *Seventeenth Annual IEEE Symposium on Semiconductor Thermal Measurement and Management*, 2001, pp. 71-84.

[11] Vuolevi, J., and T. Rahkonen, "Extraction of nonlinear AC FET model using small-signal S parameters," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 50, No. 5, May 2002, pp. 1311-1315.

[12] Vuolevi, J., J. Aikio, and T. Rahkonen, "Extraction of a polynomial LDMOS model for distortion simulations using small-signal S-parameter measurements," *2002 Microwave Theory and Techniques Symposium*, Seattle, WA, pp. 2157-2160.

[13] Lu, K., P. Perry, and T. Brazil, "A new SPICE-type heterojunction bipolar transistor model for DC, microwave small-signal and large-signal circuit simulation," *IEEE 1994 MTT-S International Microwave Symposium Digest*, 3, pp. 1579-1582.

[14] Dienot, J., et al., "A new characterization approach to extract HBT's models for non-linear microwave CAD," *IEEE 1994 MTT-S International Microwave Symposium Digest*, 2, pp. 977-980.

[15] Vuolevi, J., and T. Rahkonen, "Extracting a polynomial AC FET model with thermal couplings from S-parameter measurements," *Proc. 2001 IEEE International Symposium of Circuit and Systems*, Sydney, Australia, May 6-9, 2001, Vol. II, pp. II.461-II.464.

[16] *The Impedance Measurement Handbook*, Agilent Technologies, 2000.

[17] Sevic, J., "A sub-Ω load-pull quarter-wave prematching network based on two-tier TRL calibration," *Microwave Journal*, March 1999.

[18] Ludwig, R., and P. Bretchko, *RF Circuit Design: Theory and Applications*, Upper Saddle River, NJ: Prentice-Hall, 2000.

[19] Call, J., and W. Davis, "A large-signal scattering parameter measurements for RF power transistors," *IEEE 2000 Radio and Wireless Conference*, pp. 143-146.

[20] Butler, J., et al., "16-term error model and calibration procedure for on-wafer network analysis measurements," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 39, No. 12, 1991, pp. 2211-2217.

[21] Yanagawa, S., H. Ishihara, and M. Ohtomo, "Analytical method for determining equivalent circuit parameters of GaAs FETs," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 44 No. 10, 1996, pp. 1637-1645.

[22] Enz, C., et al., "MOS transistor modeling for RF IC design," *IEEE Journal of Solid-State Circuits*, Vol. 35, No. 2, February 2000.

[23] Tsividis, Y., *Operation and Modeling of the MOS Transistor*, New York: McGraw-Hill, 1987.

[24] *CLY 2 GaAs Power MESFET datasheet*, Infineon Technologies, 1996.

[25] Wambacq, P., and W. Sansen, *Distortion Analysis of Analog Integrated Circuits*. Norwell, MA: Kluwer, 1998.

Chapter 6

Simulating and Measuring Memory Effects

The main effort in Chapter 4 was in analyzing the electrical and electrothermal IM3 components of the amplifier as a function of tone spacing by means of third-order Volterra analysis. This viewpoint will be extended in this chapter. First, the effects of signal amplitude are taken into account in simulations of real PAs, extending the simplified analysis presented in Chapter 3. Second, a method for recognizing memory effects using conventional harmonic balance simulation is presented. Third, a technique for measuring both the amplitude and phase of IM3 components is presented.

Provided that we have a simulation model we can trust, the memory effects can be simulated using harmonic balance instead of the Volterra analysis, by sweeping both the tone spacing and signal level. A normalization method of IM responses that shows the memory effect more clearly is presented in Section 6.1. This normalization does not require any internal information about the simulation model, rather just the theoretical amplitude ratios of the fundamental, IM3, and IM5 components. Therefore, the method is not limited to polynomial Volterra models, but can be applied to any kind of nonlinear model. The results, of course, can be at most as accurate as the used simulation model, however.

Since it is unfortunately quite usual that the simulation models fail to simulate such high-level nonlinear phenomena as memory effects, it is advantageous for one to know how to measure these effects. Therefore, a measurement technique for characterizing the memory effects in a real PA is presented in Section 6.2, and the measured results for BJT and MESFET amplifiers are given to let the reader have some idea of the real importance of the memory effects. Although the conclusions for the seriousness of the

memory effects with a particular linearization technique is left to the reader, the memory effects from the linearization point of view will be briefly discussed in Section 6.3. The most important results of this chapter are summarized in Section 6.4.

6.1 Simulating Memory Effects

The third-order Volterra model is a good tool for recognizing distortion mechanisms and memory effects, but due to the third-degree modeling it cannot predict how the memory effects of IM3 vary with signal amplitude, as already briefly discussed in Section 3.5. A fifth-order Volterra model is capable of doing this, but an analytical solution of the fifth-order expansion grows far too complicated and is not presented here.

Instead of the Volterra analysis, standard RF simulators are used here for simulating the memory effects. The harmonic balance (HB) method is commonly used for nonlinear RF simulations [1]. In spite of some problems with convergence, numerical noise, and simulation speed, harmonic balance is a very useful tool when used correctly [2], and the problems are more often related to models than the algorithm itself. Most models of active and passive components are inaccurate at frequencies far away from the fundamental frequency, and therefore it is safe to use a moderate number of harmonics in the simulations. For example, if nonlinearities up to the ninth-degree in a 2-GHz amplifier are to be taken into account, the models must be valid up to the ninth harmonic at 18 GHz. Any discontinuity in derivatives or other nonphysical phenomena at that frequency will affect IM3, and consequently the amount of harmonics in HB should be chosen so that the frequency of the highest harmonic does not exceed the range in which the simulator models work adequately [3].

6.1.1 Normalization of IM3 Components

The drawback of the HB method is that it displays each spectral tone as a total result, and there is no way of seeing the fine structure of the distortion. This section presents the technique to gain some insight into the contributors of IM3 based just on the magnitude and phase results of the IM3 and IM5 tones. Although this normalization does not give the detailed fine structure of the distortion that the Volterra analysis presents, it can be used to recognize the memory effects of the circuit to be simulated.

Let us first take a look at the conventional way to plot the IM3 components as a function of signal amplitude. Logarithmic input and output amplitude axes are normally used, resulting in a line with a slope

equal to 3:1, as seen in Figure 6.1(a). Actually, Figure 6.1(a) contains two nearly overlapping curves, which start to deviate only at high amplitude levels: the upper one presents the IM3 caused by cubic nonlinearity alone, while the lower one includes both third- and fifth-degree nonlinearities. However, due to the logarithmic scales used, the difference between the two is difficult to see. The same information is presented in Figure 6.1(b) in a different way. Normalized scales are used, and the cubic amplitude dependence of IM3 is normalized by dividing the IM3 amplitude by the third power of the fundamental input amplitude. As a result, a pure cubic nonlinearity yields a constant value straight line, shown by a thick line in Figure 6.1(b). The curve, including both third- and fifth-degree nonlinearities, is plotted with a thin line in Figure 6.1(b), and the compression due to the fifth-order term is clearly visible. The nonlinearities in both figures are the same, but evidently the fifth-order effects of IM3 can be recognized more easily from the normalized plot.

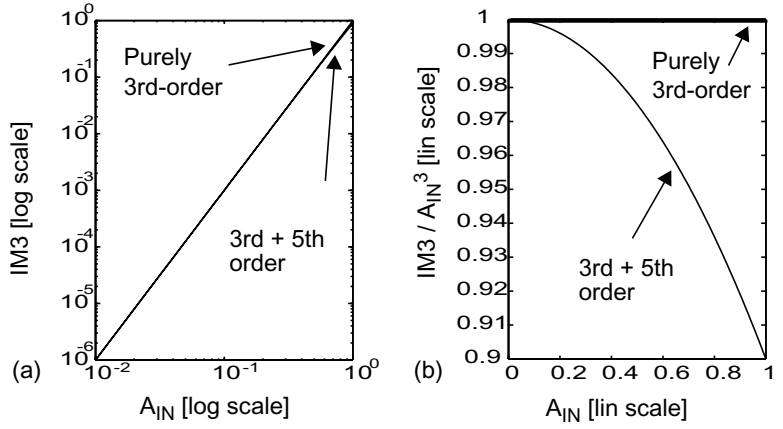


Figure 6.1 IM3 as a function of input amplitude using (a) logarithmic axis and (b) linear axis.

Once the 3:1 dependency is removed from IM3, the next step is to separate the fifth-degree nonlinearities to have a clearer look at the memory effects of the IM3 components. This is needed because the fifth-order effects are usually so strong that the memory effects would be masked by the fifth-order distortion at high amplitude levels. The normalization can be seen as a search for the coefficients a_3 and a_5 of a polynomial input-output nonlinearity. If the system does not exhibit nonlinearities higher than the fifth degree, and it does not have memory effects, the coefficients a_3 and a_5

describe the behavior of IM3 at all modulation frequencies and amplitudes. However, in practice memory effects always exist, and now the idea of the normalization is to compare the polynomial input-output estimated IM3 value to the real one, and the memory effects can be seen as a deviation between the two.

In Section 3.5.1 the spectral composition of a two-tone test in a memoryless fifth-degree nonlinearity was presented. It was noticed that the amplitudes of IM3 and IM5 caused by fifth-degree nonlinearity have factors of 25/8 and 5/8, respectively. So the a_5 term contributes to the IM3 and IM5 in different ways, and the amplitude ratio between the two is five and the phase difference is zero, provided that no memory effects exists. Since the analysis is truncated to the fifth-order, IM5 is caused by a_5 only, and the IM3 caused by a_5 can be estimated based on IM5. Now the entire normalization can be written as

$$IM3_{\mathrm{NORM}} = \frac{IM3 - 5 \cdot IM5}{A_{IN}^3} \tag{6.1}$$

The denominator in (6.1) is needed to remove the 3:1 the dependency of IM3, and since five times the IM5 is subtracted from IM3, normalized IM3 includes just the third-order distortion. Now, the system including third- and fifth-order distortion without memory effects yields normalized IM3 values that should be constants as functions of input amplitude and tone spacing. With memory effects, either the phase difference between the fifth-order IM3 and IM5 tones or their amplitude ratio varies, and this can be detected as humps or dips in the normalized IM3 plane. Therefore, a nonconstant value of normalized IM3 indicates memory effects. The above reasoning holds as long as there are no disturbances from seventh or higher order distortion: when this is no longer true, the normalized IM3 values start to deviate at high amplitude values also without the memory effects. Memory effects can then be viewed as deviations between the actual shape of IM3 and the one predicted by a polynomial input-output model.

The ratio 5 in (6.1) may be affected by higher order effects, and a more accurate guess can be obtained by a reference measurement at a presumably low-memory point in the amplitude-tone spacing plane. Here the best solution might be to pick a narrow tone spacing to avoid high-frequency memory effects and disable the thermal memory effects either by fixing the junction temperature or by setting the thermal impedance purely resistive. From this reference simulation we can pick a more accurate guess to the ratio of IM5 and IM3 caused by the fifth-order distortion, so that the normalized IM3 flattens with respect to amplitude. Then we can use this

new guess in place of the coefficient 5 in (6.1) to normalize the entire IM3 plane. Also in this case, the memory effects can be seen as deviations of the A_{IN}-IM3 curve at different tone spacing, compared to the value obtained at narrow spacing without memory effects.

6.1.2 Simulation of Normalized IM3 Components

The normalization (6.1) is now applied to the CLY2 common-source FET amplifier analyzed in Chapter 4. As already noted in the beginning of this chapter, the normalization does not require any information about the simulation model and can be done for any nonlinear model. Instead of using standard FET models, the simulation model used here is based on polynomial nonlinearities. The polynomial nonlinearity coefficients of the model are extracted up to the fifth degree by the *S*-parameter characterization method presented in Chapter 5. The simulation model of the amplifier is presented in Figure 6.2(a) and the nonlinearities of the circuit elements are modeled by polynomial voltage-controlled current sources (VCCS, available both as conductive and charge elements in the Aplac simulator [4]), corresponding to the principles of the Volterra analysis. The C_{GS} is modeled as a charge source, the value of which is a nonlinear function of v_{GS}, and can be expressed similar to (4.6). The drain-to-source current is a function of v_{GS} and v_{DS}, and it can be expressed similar to (4.3). Compared to these equations, the VCCSs used here include electrical nonlinearity coefficients up to the fifth-degree, and, for simplicity, the electrothermal nonlinearity coefficients are neglected, resulting in a purely electrical distortion simulation. The terminal impedances are measured from an existing amplifier designed according to the data sheet [5].

A two-tone test at the center frequency of 1.8 GHz is applied, and the tone spacing and the power of the input signal are swept. Without any normalization, the simulated amplitude of IM3H in decibels with respect to the fundamental (dBc) is presented in Figure 6.3, and a linearly increasing surface with respect to increasing signal amplitude is obtained. The surface is practically flat as a function of tone spacing, except for a very small variation at 500 kHz and high amplitude values. This variation at 500 kHz is caused by a low-frequency LC resonance in the drain bias circuit, and the effect will be explained in more detail in Section 6.2.4.

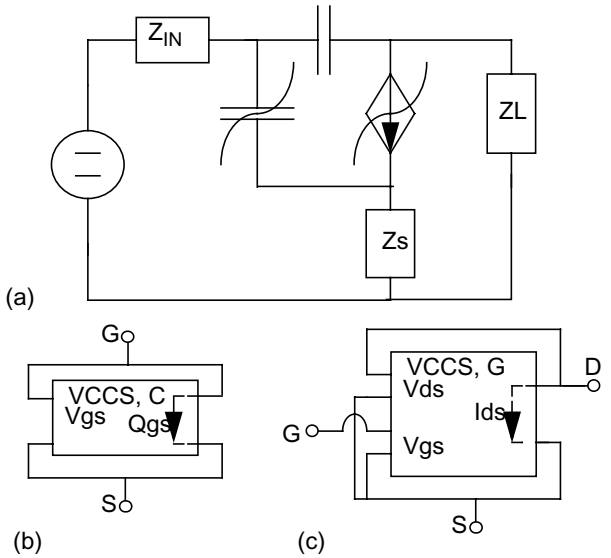


Figure 6.2 (a) Amplifier model to be simulated, (b) nonlinear capacitance, and (c) 2D-transconductance build using VCCSs.

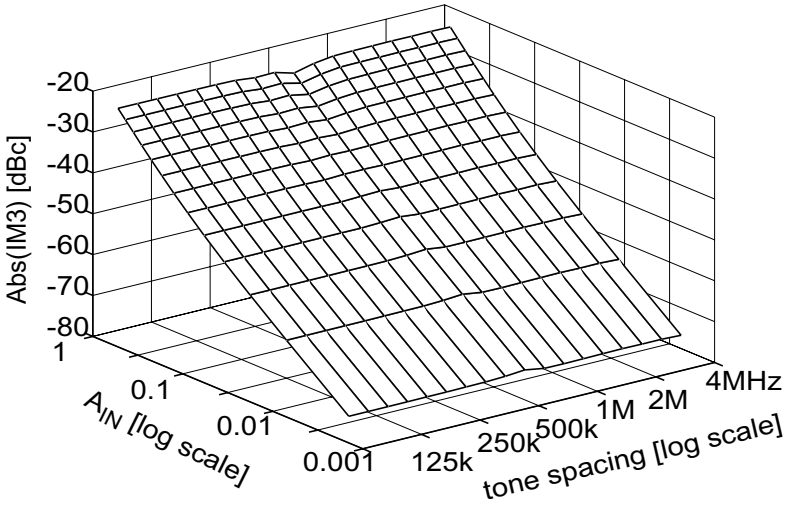


Figure 6.3 IM3 in dBc as a function of tone spacing and amplitude. Both the frequency and amplitude axis are logarithmic.

To see how IM3 deviates from the 3:1 slope, let us now apply the first part of the normalization and divide the IM3 value by a third power of the input amplitude. The magnitude and phase of the partially normalized IM3 are shown in Figure 6.4. Since the magnitude of it is still increasing with increasing amplitude, the fifth-degree nonlinearity expands the IM3 response. Some memory effects are seen at low amplitude values that were almost completely masked in the logarithmic plot in Figure 6.3. However, now the effects can easily be seen from Figure 6.4(a). At low amplitude levels the IM3 value peaks at 500 kHz, but at high amplitudes the situation is almost the opposite as the normalized amplitude dips lowest at 500 kHz. Since the shape of the memory effects varies with amplitude values, amplitude dependent memory effects clearly exist. Figure 6.4(b) presents the phase of the normalized IM3, and here, too, the memory effects at 500 kHz are seen, but since the overall behavior is still dominated by fifth-order distortion, the effects are difficult to pinpoint.

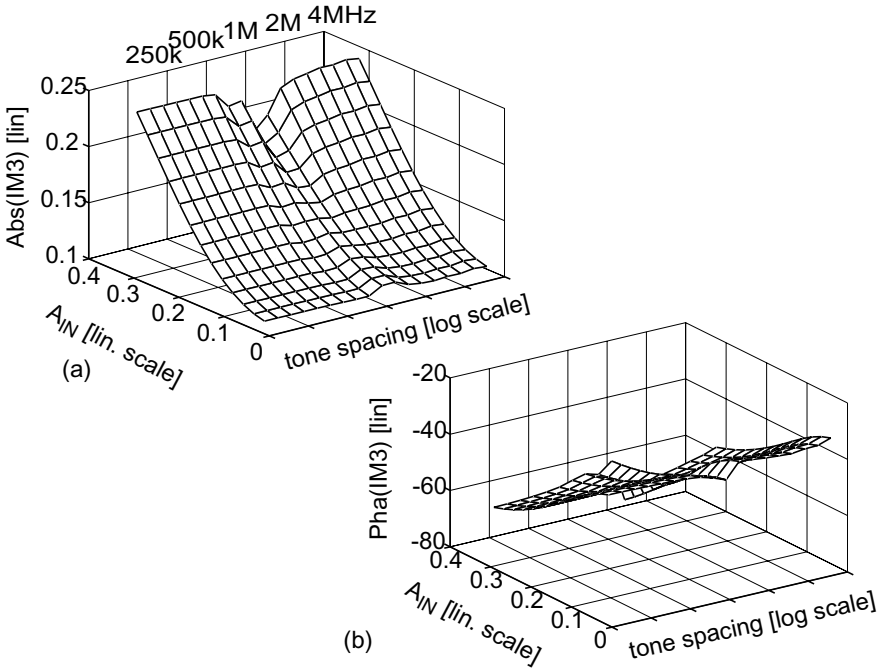


Figure 6.4 Partially normalized (a) magnitude and (b) phase of the IM3 as functions of tone spacing and amplitude. Normalization is done by dividing the value of IM3 by the third power of the input amplitude, and the result is shown on a linear scale.

It is observed by comparing Figures 6.3 and 6.4 that just by dividing the IM3 amplitude by the third power of input amplitude and using linear scales, a lot of information about memory effects that was masked by presenting the data using a logarithmic axis can be recovered. However, the study of memory effects of IM3 still suffers from fifth-order distortion that is bending the partially normalized planes at high amplitude levels. Therefore, the full normalization is now applied to remove also the fifth-order effects from IM3. To be accurate, not all the fifth-order distortion is removed by the normalization, rather only the memoryless part that can be predicted by a simple input-output polynomial model. The magnitude and phase of the normalized IM3 are presented in Figure 6.5. The amplitude dependencies are removed almost completely, indicating that the IM3 is mostly caused by third- and fifth-order distortion. Only at very high amplitude levels the seventh- and higher order effects start to affect the IM3, causing the surface to bend slightly again.

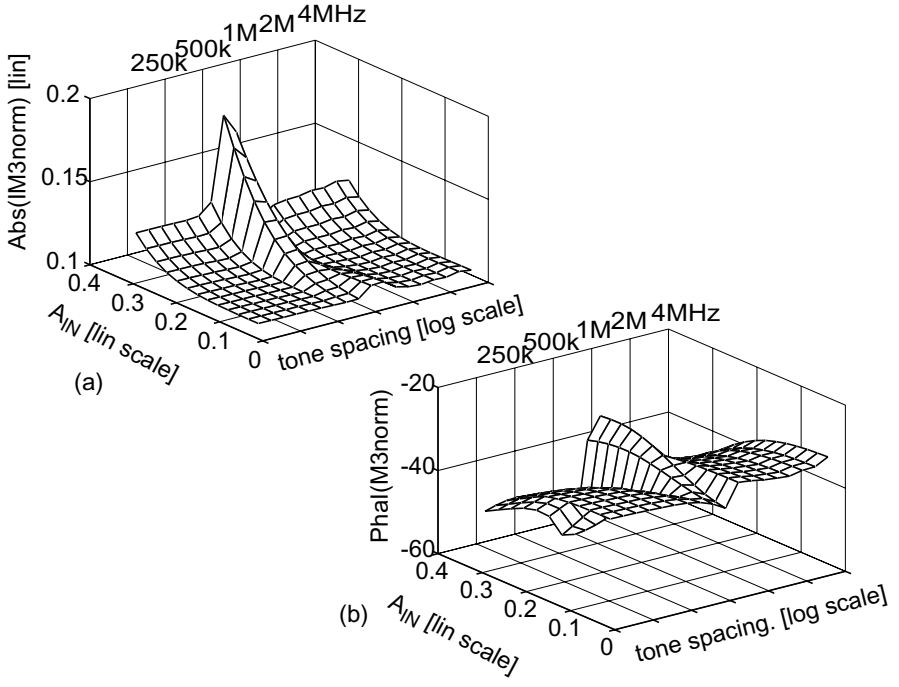


Figure 6.5 Normalized (a) magnitude and (b) phase of the IM3 as functions of tone spacing and amplitude presented in a linear scale. The IM3 is normalized according to (5.1).

It is very interesting to look at the behavior of the fully normalized IM3 around the resonance at 500 kHz in Figure 6.5. The highest peak in the normalized IM3 is at 500-kHz tone spacing at low amplitude levels, but it appears at 250 kHz at high amplitude levels. This phenomenon is observed from both the amplitude and phase surfaces and it can be explained using Figure 6.6 as follows. With the tone spacing of 500 kHz [shown in Figure 6.6(a)], the second-order envelope ω_2–ω_1 falls on top of the resonance in the load impedance and mixes strongly back to IM3, causing a bump in the response. At the tone spacing of 250 kHz [Figure 6.6(b)], it is now the fourth-order envelope component ($2\omega_2$–$2\omega_1$) that falls on the resonance at 500 kHz and upconverts further to IM3 and IM5. Since the amplitude of the fourth-order envelope component is proportional to A_{IN}^4, this effect dominates at high signal levels but vanishes at low amplitudes, leaving just the bump caused by the second-order envelope (ω_2–ω_1). Thus, the effects of the bias resonance appear at different tone spacings at different signal levels.

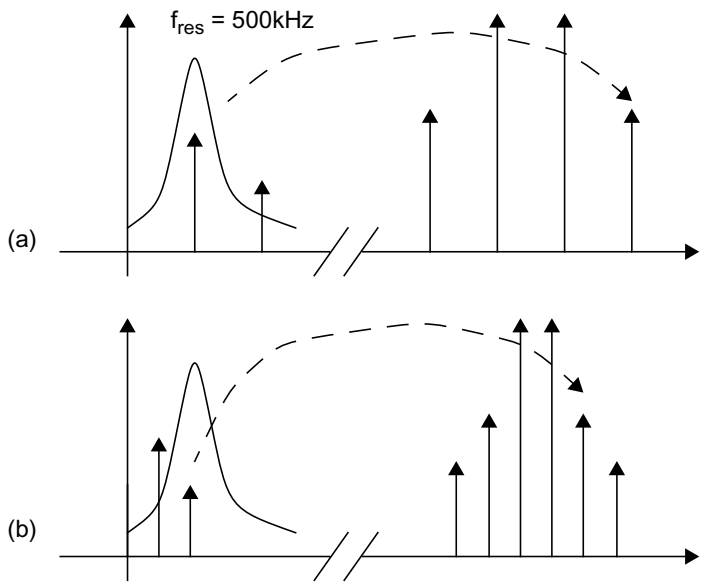


Figure 6.6 The two-tone spectrums with tone spacing (a) 500 kHz and (b) 250 kHz. The 500 kHz resonance in the collector bias impedance Z_L is plotted on top of the baseband spectrum.

6.2 Measuring the Memory Effects

This section presents a three-tone test setup capable of providing amplitude and phase information of IM3 signals. As examples, measured results of memory effects in CE BJT and CS MESFET amplifiers are presented.

Memory effects are rather difficult to measure. Spectrum analyzers may be used to measure sideband amplitudes, but they do not provide phase information. A system comprising two network analyzers (see [6]) is capable of yielding phase information on both fundamental signals in a two-tone test, but the measurements only give information on the memory effects of the fundamental signal, and can be regarded as the modulation frequency dependence of the AM-AM and AM-PM curves. Since the behavior of IM3 components differs from that of fundamental signals (for example, in third-order analysis the fundamentals are affected by the rectified DC voltage while IM3 tones are not), this procedure does not give complete information on the memory effects of the IM components, which is of primary interest in terms of linearity and linearization.

Several attempts have been made to measure the phase of intermodulation and harmonic tones. These are usually based on a high-speed diode that is used as a reference nonlinearity, producing a constant-phase IM3 component over the modulation band. Figure 6.7 shows that kind of a setup proposed in [7]. In this test setup, the input signal is split into two branches, and the first of them is connected to the DUT and another one to the diode. Both the diode and the DUT generate intermodulation tones, that are combined and connected to the spectrum analyzer. If the measured tone (IM3) disappears from the spectrum analyzer, the amplitudes of the tones generated by the DUT and reference nonlinearity are equal, and their phase difference is 180º. The setup is based on this information. However, since the characteristics of the reference nonlinearity and the nonlinearity of the DUT are different, some amplitude and phase tuning after the DUT is needed to obtain cancellation. Since the amplitude level at the input of the diode needs to be held constant, an attenuator before the DUT is needed to control its input level.

The drawback of the method is that only relative phase information can be obtained. The validity of the results depends on the reference nonlinearity and an ideal third-order distorter is needed to avoid errors. Memory effects in the reference are liable to cause errors in the results. From a practical point of view, one of the most harmful drawbacks is that the tuning of manual attenuators and phase shifters at every amplitude and tone-difference value is quite a lengthy task.

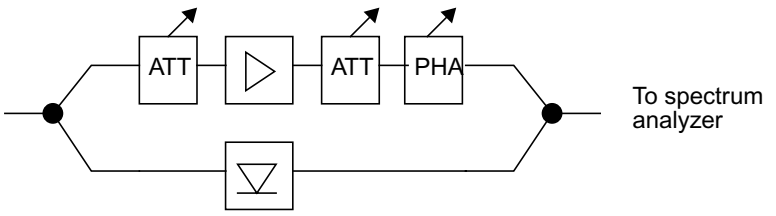


Figure 6.7 IM3 phase measurement using a reference nonlinearity. After [7].

6.2.1 Test Setup and Calibration

The test setup presented in this book is also based on the cancellation of intermodulation tones, and its operation is illustrated in Figure 6.8(a). Instead of generating the cancellation using a reference nonlinearity, a cancelling tone (A_3) is simply injected together with the fundamental two-tone signal (A_1 and A_2) to the input of the amplifier. If we are measuring the IM3L, the frequency of the injected A_3 is $2\omega_1-\omega_2$, and for measuring IM3H it is $2\omega_2-\omega_1$.

Let us now discuss the cancellation of IM3 at the output of the amplifier. If we assume the amplifier to be a third-degree polynomial, its IM3L phasor is $(3/4)a_3A_1^2\overline{A}_2$. However, at the output of the amplifier, the A_3 produces a tone at the same frequency, which can be written as a_1A_3. If we now tune the amplitude and phase of A_3, we come to a situation in which the IM3 component at the output of the amplifier disappears. This is basically the actual measurement, but some care is required to find out the phase of the signal generator A_3.

If we perform a tone spacing sweep of the two-tone signal, we naturally have to change the frequencies of the signal generators. Here we lose our phase information. Some readers may doubt the relevance of phase information of the two-tone signal and they are right in the sense that the phase difference between two signals at different frequencies varies with time. This is usually not an important consideration in RF design. However, the test setup here is a three-tone test setup and the tone difference between A_1 and A_2 equals the tone difference of A_3 and A_1 (or A_2 and A_3 in case of IM3H), and in this case, the phases of the signals are significant. The calibration of the phase is explained in Figure 6.8(b). The mixing of the

tones of the two-tone signal causes an envelope component (ω_1–ω_2), which can be written as $a_2\overline{A}_1A_2$. Similarly mixing the A_1 and A_3 causes a component also at the envelope frequency, the amplitude of which is $a_2\overline{A}_1A_3$. If we sum up these two envelope components together and adjust the amplitude and phase of A_3, a situation where the total envelope component vanishes can be obtained. This gives us a phase reference for A_3 and this must always be repeated when the frequencies of the signal generators are changed.

Figure 6.9 presents the practical test setup. Power splitters are used to divide all three tones into two branches, and the upper branches (marked by the number 1 in Figure 6.9) are combined together and applied to the DUT to produce the required three-tone signal. The actual test signal is therefore a sum of ω_1, ω_2, and $2\omega_1$–ω_2, all locked to the same reference, and the rest of the circuit is needed for calibration. The lower branches of the 180º

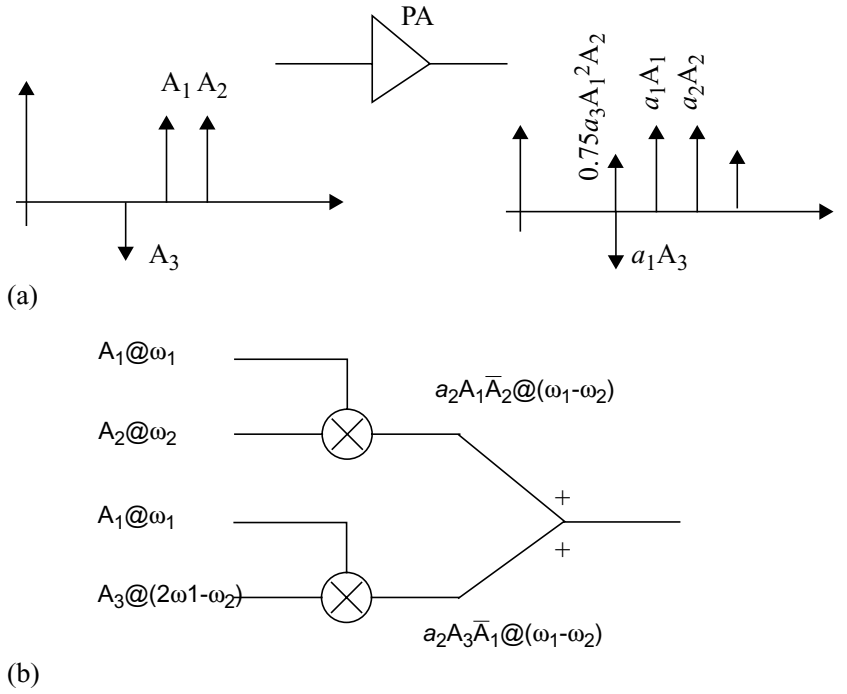


Figure 6.8 (a) Cancellation using IM3 tone injection, and (b) phase calibration of the test setup.

power splitters (marked by the number 2 in Figure 6.9) are used for mixing the phase reference, which is done by comparing the two downconverted envelope signals. These envelope signals are generated by mixing the tones of a two-tone signal, generated by signal generators ω_1 and ω_2 and mixing the IM3 signal and the lower two-tone signal, generated by the signal generators ω_1 and $2\omega_1-\omega_2$. Since the signal generator at ω_1 is needed for both envelope mixings, one more power splitter is required (in this case, a 90º splitter) and to make the test setup symmetrical, a similar power splitter is also added for ω_2, where one output terminal is connected to a 50-ohm termination impedance.

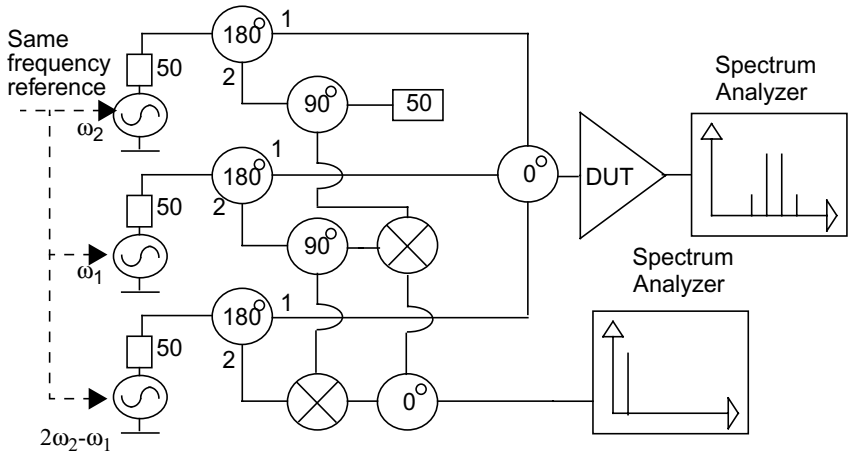


Figure 6.9 IM3 injection system for measuring memory effects. © IEEE 2001 [8].

For phase measurement, the two envelope signals are brought to a resistive power combiner, and by adjusting the amplitude and phase of the IM3L signal generator, the signal at the output of the power combiner is made to vanish. After that, the amplitude and phase of the IM3L signal are adjusted again until the IM3L component at the output of the amplifier disappears. In this way, the phase difference between these two situations in which the signal component vanishes, gives the phase of the IM3L component of the amplifier. Calibration of the cables is important in phase measurements, because the electrical length of the cables is a function of the tone spacing in the two-tone signal, and this effect has to be taken into account if the maximum modulation frequency is in the megahertz range. The low-frequency reference part of the circuit must be calibrated, too, because its electrical length also becomes important in the megahertz

range. Once the phase information is obtained, the amplitude of IM3 is very easy to obtain, being a conventional spectrum analyzer measurement. Only the attenuation of the cables and power splitters/combiners needs to be taken into account.The addition of a 90º power splitter for ω_2 also makes the test setup easily extendable. A fourth signal generator can be added at the $2\omega_2-\omega_1$ to measure both IM3L and IM3H simultaneously and an additional envelope reference is produced by mixing $2\omega_2-\omega_1$ and ω_2 down. However, this is not discussed in this book, because the sidebands can be measured independently by changing the IM3L signal generator to the IM3H frequency and by changing the frequencies of the lower and upper two-tone signals. Finally, the test setup is fully automatic, controlled by LabVIEW software [9].

It is also important to emphasize that the test setup actually does not measure the IM3 component at the output. Instead, it measures the optimum input predistortion signal that creates maximum cancellation of the output. This is a significant advantage, because it allows the measurements to be used directly as required characteristics for the predistortion circuits.

6.2.2 Measurement Accuracy

The accuracy of the measurements is an important consideration. There are two kinds of measurement error: those that cannot be calibrated and those that can be taken into account by careful calibration. The errors that cannot be corrected unless the measurement setup is improved are determined by the canceled IM3 amplitude. For example, if an IM3 level of –40 dBc is detected at the output of the amplifier and can be canceled to –70 dBc, this 30-dB cancellation performance requires the amplitude or phase error to be less than 0.25 dB or 1º. The canceled IM3 level is usually limited by the noise floor of the spectrum analyzer or phase noise of the signal generator. As a result, high performance measurement equipment has to be used to reduce the errors significantly. Unfortunately, errors of more than 1º were observed, and calibration was required to improve the performance of the test setup. It was found that the phase of some inexpensive RF signal generators is a slight function of the amplitude and that more serious phase jumps occur at fixed amplitude steps when switching the attenuators to the RF path inside the signal generator. Nevertheless, these phase errors can be calibrated, and their role can be diminished by using a high-performance RF signal generator. Furthermore, some signal generators suffer from a very slow phase drift. The effects of this defect can be avoided by rechecking the canceling phases at the end of each measurement. Also, any test setup will exhibit a degree of unwanted intermodulation, so that even if

a signal generator is not applied at IM3, signal components can still be observed there. This may turn out to be a real problem when measuring highly linear stages, but fortunately the test setup itself could be calibrated by measuring it without the DUT. The level of unwanted intermodulation responses was found to be –70 dBc, but the use of high quality mixers, power combiners, power splitters, and cables can easily reduce the figure by more than 10 dB.

6.2.3 Memory Effects in a BJT PA

This section presents the measured results of memory effects of the CE BJT amplifier, and the next section presents the measured results of the CS MESFET amplifier. The purpose of these measurements is to show the kind of memory effects and how strong they really are in the two power amplifiers, and to demonstrate how accurately the measurement test setup is capable of measuring it.

The first measured amplifier is based on a Phillips BFG 11 BJT transistor, used in a common-emitter configuration. It is a single stage amplifier that corresponds to the one presented in the data sheet provided by the manufacturer [10]. The V_{CE} and V_{BE} bias values of 3 V and 740 mV and the center frequency of 1.8 GHz are used in measurements.

The tone spacing of the two-tone input is swept, and the injected IM3 signals are tuned to achieve 25-dB cancellation in the output IM3 level. The amplitudes (in dBc) and phases of the injected IM3 canceling tones are shown in Figure 6.10 over the tone spacing range of 32 kHz and 32 MHz. The amplitude of the required canceling signal at the input is directly proportional to the IM3 level at the output. The phase of IM3 is interesting, however, as the phases of the two tones are equal at high modulation frequencies (above 500 kHz) but start to deviate at low modulation frequencies, so that a large 20º phase offset is detected at 32 kHz. Since no electrical time constant of that size exists in the circuit, these low frequency memory effects are caused by thermal power feedback.

A phase jump is seen in the response at 1 MHz. It can be explained by the schematic diagram of the amplifier presented in Figure 6.11(a), and the collector impedance at the envelope frequency given in Figure 6.11(b). The collector impedance resonates badly at 1 MHz, and this corresponds exactly to the phase jump in the tone spacing response. Evidently, the phase jump is caused by the collector bias circuit resonating at the envelope frequency.

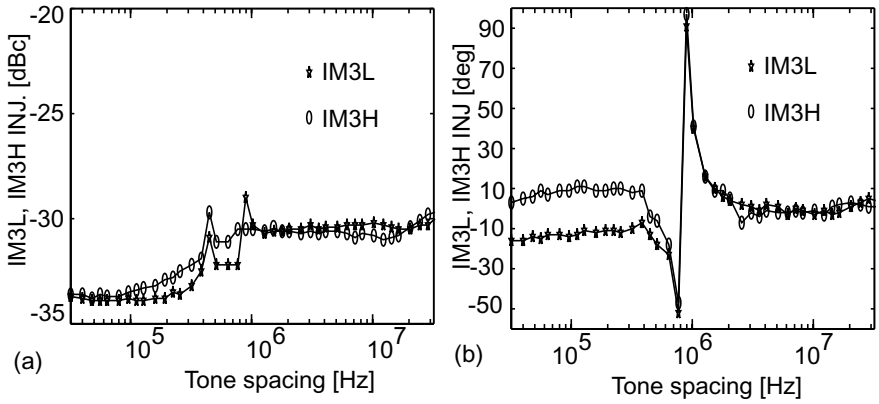


Figure 6.10 Measured (a) amplitude and (b) phase of optimum injected predistortion signals for a 25-dB cancellation in a Philips BFG 11 common-emitter amplifier over the range of modulation frequencies at a constant fundamental input amplitude. © IEEE 2001 [8].

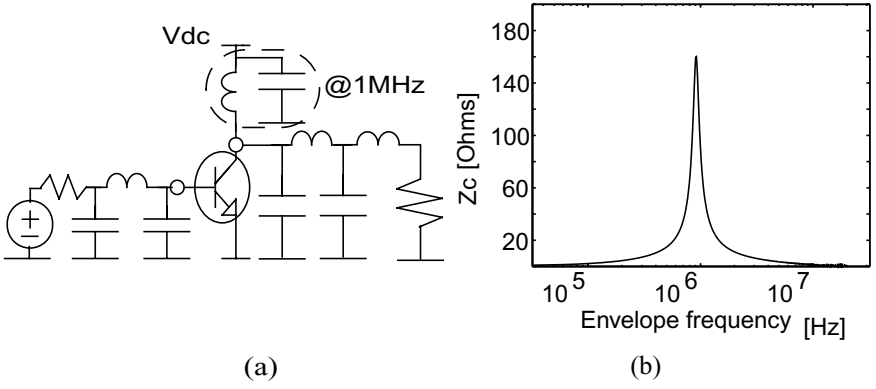


Figure 6.11 (a) Schematic diagram of the BJT amplifier, and (b) measured impedance of the collector node as a function of modulation frequency. © IEEE 2001 [8].

6.2.4 Memory Effects in an MESFET PA

Now the measured results of the Infineon CLY2 amplifier [5] are presented. Drain bias voltage and current of 3V and 20 mA, respectively, and the center frequency of 1.8 GHz are used in the measurement.

Again, a tone-spacing sweep is performed, and the injection IM3 signals are tuned to achieve optimum cancellation. The amplitudes of the required injected canceling signals remain practically constant over the entire modulation band, as seen in Figure 6.12(a), but this does not mean that no memory effects occur. Instead, the phases of the cancelling IM3 tones start to deviate above 1 MHz, and a phase difference of 40º is met at 30 MHz. In addition, a smooth phase bump of 10º around 500 kHz can be seen. In this case, no thermal memory effects are detected at low modulation frequencies, and this is mainly because the MESFET is biased to be quite nonlinear, resulting in a predominance of the purely electrical causes of distortion.

Since it was shown in Chapters 3 and 4 that the majority of electrical memory effects are caused by the impedances at the envelope frequency, the baseband gate and drain node impedances of the amplifier are plotted in Figure 6.13. The drain impedance is fairly constant over the modulation band, but the value at 500 kHz is relatively high, caused again by an LC resonance in the dc bias feed. This resonance causes a bump of 10º in the phase of the IM3 responses. At the gate side, resistive gate biasing is used to reduce memory effects at the input, but still the gate impedance starts to decrease at high modulation frequencies due to the C_{GS}. As a result, the phases of the upper and lower IM3 tones start to diverge from one another.

6.3 Memory Effects and Linearization

Throughout this book the decreased performance of linearization caused by memory effects is pointed out. Although more detailed linearization measurements will be presented in Chapter 7, here the measured memory effects of the amplifier and cancellation performance of analog predistortion are presented side by side. A polynomial RF predistorter is used to cancel the IM3 components of the MESFET amplifier, and quite good cancellation of 22 dB is obtained at narrow tone-spacings, as seen in Figure 6.12(a). At the resonance frequency of 500 kHz the cancellation drops to 12 dB, but it returns to 20 dB at frequencies above that. However, when the tone spacing is increased beyond 2 MHz, the cancellation performance drops completely. The comparison of IM3 phases and the cancellation performance indicates that the cancellation performance

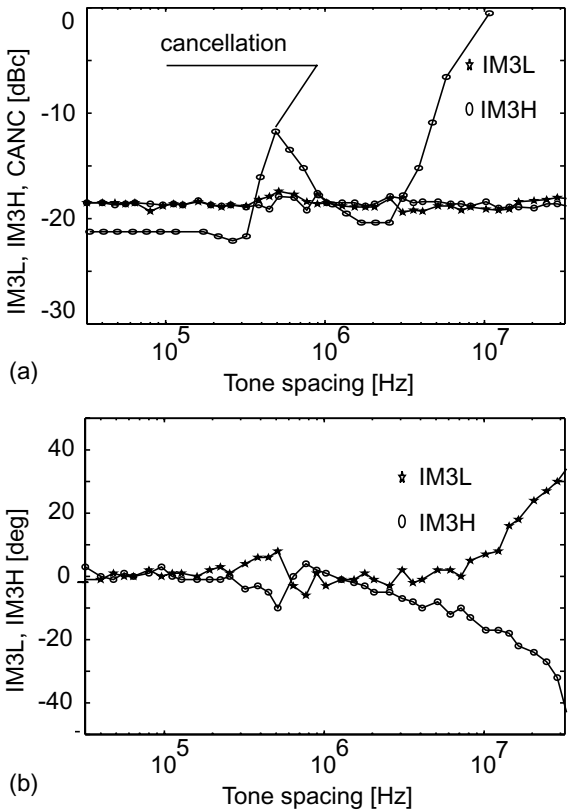


Figure 6.12 (a) Amplitude and (b) phase of the optimum predistortion signals measured for the Infineon CLY 2 common-source amplifier over the range of modulation frequency at a constant fundamental input amplitude. Plot (a) also shows the cancellation achieved using polynomial RF predistortion. © IEEE 2001 [8].

strongly correlates with the phase tracking of the IM3 tones. Just a weak resonance, masked completely in amplitude measurements of IM3, and observed just as 10º of phase variation, can reduce the achievable cancellation from 22 dB to 12 dB, as predicted in Figure 3.3 in Chapter 3.

The above study shows that due to the memory effects, a linearizer can be tuned for a good cancellation of distortion at one specific tone spacing (0 Hz in the previous example) and one amplitude, but if either one changes, the cancellation performance will deteriorate. This problem is

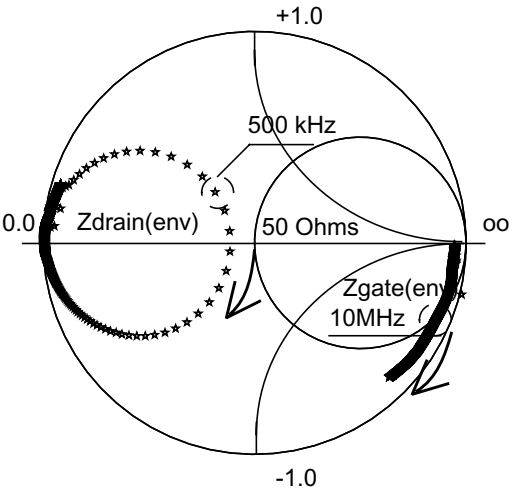


Figure 6.13 Measured gate and drain node impedance of the CLY 2 amplifier from 32 kHz to 32 MHz. © IEEE 2001 [8].

usually handled by using more sophisticated and power consuming linearization methods that are capable of adapting to varying signal conditions. If simple linearizers like RF predistorters are used, there are two approaches to overcoming the problems of memory effects. First, the opposite memory effects can be constructed inside the predistortion circuit, or second, the memory effects of the PA must be minimized.

The normalized IM3 surfaces give a clear picture of how easily the power amplifier can be linearized. A flat surface indicates that the amplifier behaves like a memoryless input-output polynomial, which is optimal in terms of linearization. Therefore, the optimum cancellation performance over the ranges of both the amplitude and modulation frequency can be achieved if the normalized IM3 surface is flat. The normalized IM3 surface is very useful for predicting the cancellation performance and/or adaptation requirements of a selected linearization technique. Moreover, the technique can be employed for optimizing the performance of the power amplifier. The ease of linearization of an amplifier can be improved by using carefully selected matching impedances at different frequency bands, for example. This new figure of merit, called here the *linearizability*, describes how easily the power amplifier can be linearized, and it is very important for the design of power amplifiers for modern telecommunication systems. In many cases, poor cancellation performance of a predistorter, for

example, is a result of an improperly designed power amplifier rather than any fault in the linearizer.

Very few papers so far have dealt with improving the linearizability of the amplifiers. Almost all the technical and scientific papers in the field describe just how much cancellation the linearizers concerned are able to produce, but in reality this figure is closely related to the linearizability of the amplifier. Some work has been done to study how the conduction angle and operating class affect the performance of analog predistortion, for example, (see [11]), but as seen above, the linearizability is a more complicated phenomenon.

6.4 Summary

Memory effects have to be taken into consideration in linearized power amplifiers. Bandwidth-dependent memory effects can rather easily be calculated analytically by means of a third-order Volterra model, but the amplitude domain effects necessarily need numerical simulations. Since numerical tools such as HB are only able to show the sum of each nonlinear response, a normalization technique is introduced to separate the fifth-order distortion from IM3. The ratio between the IM3 and IM5 components of the fifth-order distortion is known, provided that no memory effects exist and that higher than fifth-order distortion is negligible. Based on this information, the fifth-order distortion can be separated from IM3, and a constant normalized IM3 value as a function of amplitude is obtained. If the amplifier has memory effects, the separation is no longer perfect, and memory effects can be identified from the normalized value of IM3. Both the tone spacing and signal amplitude can be swept and the values at which the memory effects become significant can be seen. If the normalized surface deviates from a flat one, the amplifier exhibits memory effects and behaves differently to a single nonlinearity described by an input-output polynomial.

An interesting phenomenon in amplitude domain memory effects was noted in this chapter. The resonance in the drain bias impedance causes memory effects at the resonance frequency of 500 kHz at low amplitude values, but at high amplitude values the memory effects were moved to the tone spacing of 250 kHz. This can be explained by the fact that the fourth-order envelope component ($2\omega_2$–$2\omega_1$) lies at 500 kHz for a tone spacing of 250 kHz, and the memory effects of the fourth-order envelope further upconverts to the IM3.

Memory effects can also be characterized by measurements. The presented three-tone test setup provides not only amplitude, but also phase

information of the distortion sidebands. The measurements demonstrate the effects of input and load impedance at the envelope frequency on the modulation response. We conclude that these impedances have to be designed very carefully when designing easily linearizable power amplifiers and the measurement data provides information that can be used to design optimal amplifiers in terms of memory effects.

6.5 Key Points to Remember

1. The term linearizability is introduced to describe how well the power amplifier can be modeled by a memoryless polynomial model, which is optimal in terms of linearization.
2. Normalization of the IM3 amplitude can be applied to find out how much the actual amplifier deviates from a simple input-output polynomial that does not have memory effects.
3. Flat surface of normalized IM3 with respect to modulation frequency and amplitude corresponds to the optimum linearizability.
4. The bandwidth of the distortion increases with signal level and the order of distortion. A resonating bias impedances causes memory effects when the tone spacing equals the resonance frequency at a low signal level, but when the amplitude is increased and fourth-order distortion dominates the baseband behavior, the effect appears at a tone-spacing that is half of the low-amplitude value. To avoid the memory effects caused by the fourth-order envelope signal the dc bias impedances must be flat up to twice the signal bandwidth.
5. The amplitude and phase of the IM3 components can be measured using a three-tone test setup.
6. Thermal memory effects caused up to 20º of phase difference between IM3L and IM3H in the measured CE BJT amplifier.
7. Resonance in IM3 responses caused by resonating collector/drain bias impedances was detected in both the CE BJT and CS MESFET amplifier.

References

[1] Kundert, K., and A. Sangiovanni-Vincenteli, *Steady-State Methods for Simulating Analog and Microwave Circuits,* Norwell, MA: Kluwer, 1990.

[2] *Microwave Office User's Manual II*, Applied Wave Research, Inc., 2000.

[3] Maas, S., "How to model intermodulation distortion," *1991 IEEE MTT-S International Microwave Symposium Digest*, Vol. 1, pp. 149-151.

[4] *Aplac User's Manual*, Aplac Solutions Corp., 2002.

[5] *CLY 2 GaAs Power MESFET Datasheet*, Infineon Technologies, 1996.

[6] Bösch, W., and G. Gatti, "Measurement and simulation of memory effects in predistortion linearizers," *IEEE Trans. on Microwave Theory and Techniques,* Vol. 37, No. 12, 1989, pp. 1885-1890.

[7] Suematsu, N., et al., "Transfer characteristics of IM3 relative phase for a GaAs FET amplifier," *1997 IEEE MTT-S International Microwave Symposium Digest*, Vol. 2, pp. 901-904.

[8] Vuolevi, J., J. Manninen, and T. Rahkonen, "Measurement technique for characterizing memory effects in RF power amplifiers," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 49, No. 8, August 2001, pp. 1383-1389.

[9] *Getting Started with LabVIEW*, National Instruments Corp., 2001.

[10] *BGF 11/X NPN 2 GHz RF Power Transistor Datasheet*, Philips Semiconductors 1995.

[11] Rahkonen, T. et al.,"Using analog predistortion for linearizing class A - C power amplifiers," *Kluwer Academics Journal on Analog Integrated Circuits and Signal Processing*, Vol. 22, No. 1, January 2000, pp. 31-40.

Chapter 7

Cancellation of Memory Effects

Previous chapters of this book have concentrated on understanding the memory effects, using idealized block models, Volterra models, simulation, and measurement techniques. In this chapter, various techniques to cancel the memory effects are studied.

For predistortion type of linearizers, memory effects may cause a significant decrease in cancellation performance. This problem is usually handled by using more complex linearization techniques, but in this chapter an attempt is made to overcome the problem by first canceling the memory effects and then linearizing the power amplifier by means of a simple, memoryless polynomial RF predistorter. Often predistorters do not give much cancellation of IM3 in the case of wideband, dynamic signals, but a significant amount of improvement in cancellation can be expected by minimizing or canceling the memory effects. Therefore, the performance of predistortion can be used as a figure of merit, on how accurately the memory effects can be canceled.

Three techniques are presented and studied in this chapter: envelope filtering, impedance optimization, and envelope injection. The first one, envelope filtering, is not actually a method for minimizing the memory effects of the power amplifier, but instead, opposite memory effects are built inside the predistorter. Next, the impedance optimization technique is presented. Since the electrical memory effects are caused by varying terminal impedances inside the frequency bands, the flattening of the impedances will reduce the memory effects. A source-pull measurement technique is developed to optimize the input impedance at the baseband frequencies and the impact of the envelope impedance on IM3 and memory effects will be seen. Third, the envelope injection technique is presented. It can be considered a real-time source-pull, which virtually generates optimal envelope impedances using an envelope signal. The envelope

signal is formed by squaring the RF signal, and it is summed directly to the input of the amplifier. The elegance of the method is that only the part of the distortion which exhibits memory effects is affected, and as a result the accuracy requirements for the injection signal are quite loose.

The techniques presented in this chapter are studied by various analysis tools. The Volterra analysis is applied to envelope filtering and injection techniques, while envelope injection is also analyzed by simulating the fifth-degree polynomial model described in Section 6.1.2. Furthermore, all the techniques are verified by measurements, and a few words about practical issues of the test setups are summarized in Appendix D. The envelope filtering and impedance optimization techniques are demonstrated using a CE BJT amplifier, while the envelope injection technique is studied using both CE BJT and CS MESFET amplifiers as examples.

7.1 Envelope Filtering

The envelope filtering technique does not change in any way the memory effects of the amplifier. Instead, its idea is to build opposite memory effects inside a polynomial predistorter, as a result of which the predistorter – power amplifier pair does not show any memory effects, even if both exhibit them. This technique nicely introduces the idea of canceling the memory effects, and that is why it is briefly presented here. More general and sophisticated memory effects cancellation techniques will be introduced later in this chapter.

The block diagram of predistortion with envelope filtering is presented in Figure 7.1. Its operation principle is the following. The first mixer, M1, rectifies the envelope signal by squaring the original two-tone input. The second harmonic is removed, and the envelope signal is then mixed with the fundamental tones in M2 to produce the IM3 predistortion signal. After

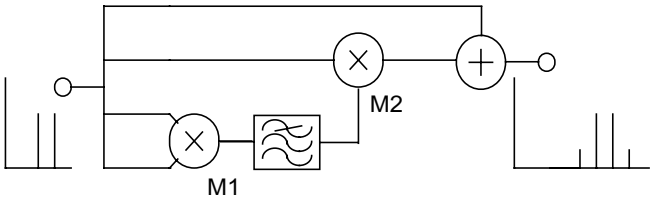


Figure 7.1 The principle of the envelope filtering technique. © IEEE 2001 [1].

tuning the phase and amplitude of the predistortion signal it is summed with the linear term in the output combiner.

In the envelope filtering technique the filter after M1 is used not only for removing the second harmonics, but also to shape the phase and amplitude response of the baseband envelope signal. As the envelope signal is a real baseband signal, its amplitude response is the same for positive and negative frequencies, but the phase response has an odd symmetry (i.e., the phase of negative frequency components is opposite to the phase of the positive frequencies). Hence, after upconversion of the envelope signal, the baseband filtering causes an opposite phase shift in the IM3 sidebands, so that if the upper sideband is leading, the lower one is lagging in phase, or vice versa. This phase difference depends on the distance to the center frequency, and the idea of this filtering is to imitate the memory effects appearing in the amplifier and maintain a 180º phase difference between the predistortion and distortion of the PA over the entire signal bandwidth.

The effects of envelope filtering are demonstrated in Figure 7.2, where the phase of the lower and upper IM3 tones are plotted as functions of the tone spacing. If the phase of the IM3 of the amplifier behaves as presented in Figure 7.2(a), a lot of improvement can be achieved by shaping the phase of the predistorter according to Figure 7.2(b). The figure depicts phase correction, but the amplitude can also be corrected in the same way and equal amplitude changes can be produced for both sidebands.

Let us now take a look at this symmetry issue in more detail. In short, the term symmetry means here that the amplitudes of the sidebands are

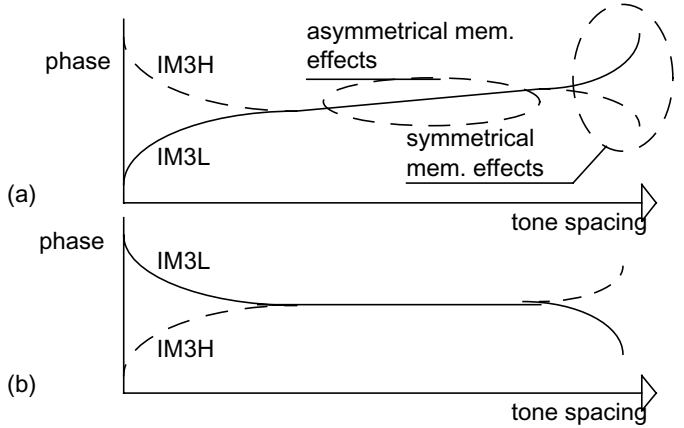


Figure 7.2 Phase of the IM3 tones (a) in the power amplifier, and (b) in a polynomial predistorter with envelope filtering. © IEEE 2001 [1].

equal but the phase shifts are opposite. As noted above, the predistorter generates IM3 sidebands with equal amplitudes and odd phase symmetry, as illustrated in Figure 7.2(b). However, the IM3 sidebands of the amplifier may have different amplitudes, and they may also have some equal-sign phase shift that cannot be corrected with the predistorter.

The feasibility of envelope filtering from a symmetry point of view is demonstrated in Figure 7.3, which shows the IM3L and IM3H components in real-imaginary coordinates. The thick lines represent the part of the IM3 components that do not show memory effects (typically caused by cubic nonlinearities), while the thin arrows represent the part of IM3L and IM3H that as a result of memory effects vary with the tone spacing.

Since the envelope filtering is able to compensate symmetrical memory effects of the amplifier, it works well with amplifiers that behave as in Figure 7.3(a), showing opposite, frequency-dependent phase deviations. The case in Figure 7.3(b) can also be corrected with envelope filtering, because it shows an equal amount of amplitude changes as a function of tone spacing. Unfortunately, problems arise when trying to correct the behavior in Figure 7.3(c) or Figure 7.3(d). Opposite amplitude deviations in Figure 7.3(c), or "common-mode" phase deviations in Figure 7.3(d) between sidebands cannot be corrected by envelope filtering, which limits the feasibility of it in the correction of amplifiers with asymmetrical memory effects.

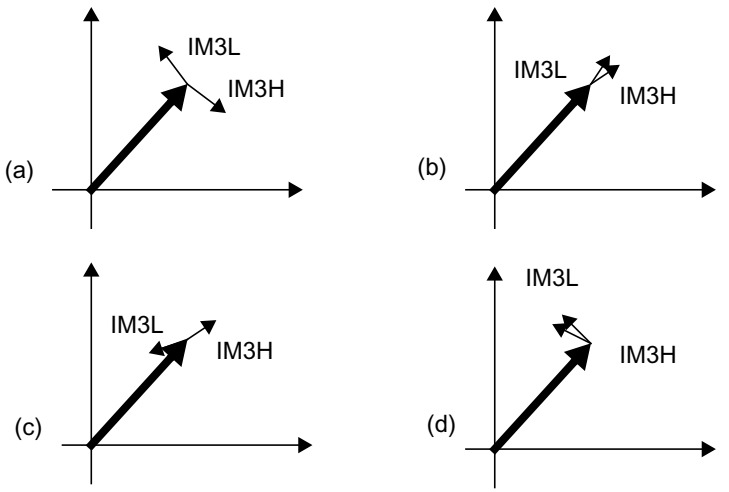


Figure 7.3 Symmetrical (a) phase and (b) amplitude memory effects. Asymmetrical (c) amplitude and (d) phase memory effects.

In order not to leave the reader a too-simplified image of this symmetry issue, one more thing needs to be discussed. The symmetry also depends on the tuning of the predistorter. If the predistorter signal aligns with the memoryless portion but has a wrong amplitude, it can still be corrected by the response of the envelope filter. However, the symmetry of cases (a) and (b) disappears if the predistorter is originally tuned so that it does not have the same phase as the thick, memoryless IM3 contribution. Therefore, it is important to tune the predistorter initially as well as possible to cancel the memoryless portion of the distortion, and then to minimize the memory effects by envelope filtering, for example.

The envelope filtering technique is tested with the same CE BJT amplifier studied in Chapter 4. Figure 7.4(a) presents the amplitudes of Volterra simulated IM3 tones in the CE BJT amplifier as a function of tone spacing. The upper curves in Figure 7.4(a) present the IM3 sidebands of the amplifier without any linearization, while the middle curves represent the sidebands linearized using a memoryless third-order predistortion. The cancellation performance of the system is limited to 15 dB by low-frequency thermal memory effects, but most of these effects can be canceled out by properly shaping the upconverted envelope signal. As indicated by the lower curves, the cancellation performance increases to 25 dB by using optimum envelope filtering.

The achieved 25-dB cancellation of the IM3 tones shows that the memory effects are very symmetrical in this case. However, the achieved cancellation is in practice determined by the accuracy of amplitude and phase responses of the envelope filtering, as well as on the accuracy of tuning of the predistorter. To check the accuracy requirements, the amplitude and phase of the envelope signal are mistuned one by one in simulations to increase the IM3 sidebands to –50 dBc, corresponding to maximum 18-dB cancellation. It is seen from Figure 7.4(b) that at the most, 2 dB of amplitude error and 10º of phase error are tolerated, if a –50 dBc IM3 level is desired. These limits arise from the cancellation accuracy discussed in Section 3.2., and as the required cancellation is nearly independent of the tone spacing, the windows for amplitude and phase errors also remain nearly constant, increasing only slightly above 1 MHz, where the required cancellation is 1 to 2 dB smaller.

From Figure 7.4(b), the shape of the required envelope filter can be reconstructed. It should have quite a flat frequency response and it should provide a 20º to 30º phase lead between 10 kHz and 1 MHz.

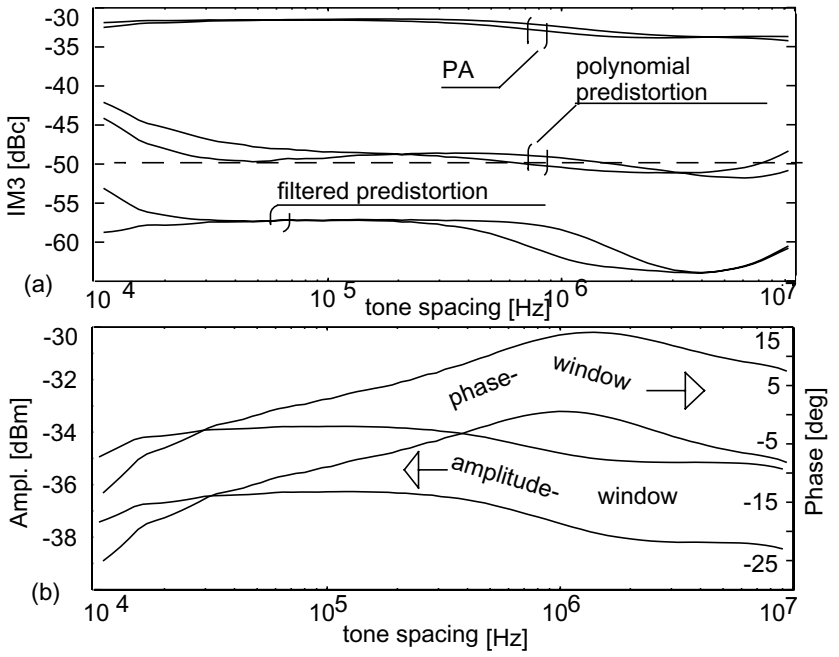


Figure 7.4 (a) Volterra-simulated IM3 improvements in memory effect cancellation, and (b) accuracy requirements in envelope filtering for an IM3 level of –50 dBc. © IEEE 2001 [1].

7.2 Impedance Optimization

As pointed out many times in this book, the electrical memory effects are caused by varying node impedances inside the frequency bands. Fundamental and harmonic bands can quite easily be designed so flat that no significant amount of memory effects arises from these frequencies. However, the same conclusion cannot be made for the envelope band, which may extend up to tens of megahertz.

The general procedure for optimizing the matching impedances of different bands is such that the fundamental input and output matching impedances are mainly optimized by the desired power, efficiency, and linearity properties of the amplifier. The harmonic impedances can be optimized for slight improvement of the efficiency as suggested in [2], provided that no narrowband harmonic traps are used, because these traps

are significant sources of memory effects and may cause large channel-to-channel variations in the amount of memory effects. In this section, the optimization of the impedances of the envelope band (i.e., the dc bias impedances) is studied. From a distortion point of view, the base impedance Z_{BB} determines how the second-order distortion currents (mostly due to base-emitter nonlinearities) entering the base terminal further transfer to distortion voltages. The nonlinear base current at the envelope frequency, for example, converts to envelope voltage in Z_{BB} and further upconverts to IM3. Therefore, it is evident that $Z_{BB}(\text{env})$ affects IM3 and memory effects. If no linearization is applied, the envelope band can be optimized to maximize the linearity, but the same procedure is no longer optimal if predistortion is used. Since the cancellation performance of linearization can be greatly reduced by the memory effects, the envelope impedance must be optimized to minimize the memory effects to get the full benefit of the linearization.

An active load principle that will virtually generate the desired impedances at the envelope frequency is presented in Section 7.2.1. The base node impedance at the envelope frequency $Z_{BB}(\text{env})$ is modified by the test setup presented in Section 7.2.2, and its effect on linearity at a constant tone spacing is monitored. Optimal $Z_{BB}(\text{env})$ contours without a predistorter are given in Section 7.2.3 and with predistorter in Section 7.2.4. The stability constraints are also discussed briefly in Section 7.2.4.

7.2.1 Active Load Principle

The active load principle is the most practical way to optimize the out-of-band impedances without affecting the fundamental matching, as the fundamental impedance cannot be kept constant by means of passive tuning. The impedance seen by a distortion tone can be affected by adding an external signal source at the same frequency. Now the apparent impedance can easily be tuned by adjusting the amplitude and phase of the signal source.

Figure 7.5 demonstrates the active load principle. The circuit can be considered a part of the CE BJT amplifier, consisting just of the input impedance Z_{IN} and the base-emitter impedance Z_{pi}. According to the notations used in Figure 3.4, the total base impedance Z_{BB} can be seen as a parallel connection of these two. Norton's equivalent of the circuit is presented in Figure 7.5(b) and a circuit that also includes the nonlinear current source of the base-emitter nonlinearities is shown in Figure 7.5(c). By applying a sum of a two-tone signal and the low frequency envelope signal to the input, the envelope voltage V_B at the base can be written as

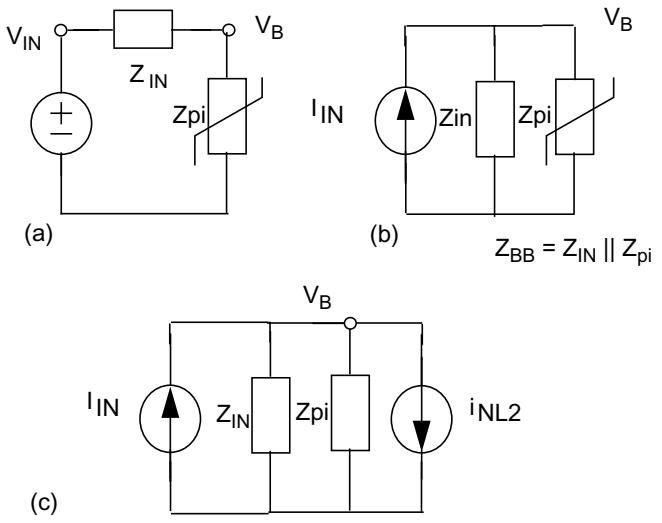


Figure 7.5 (a) A simple nonlinear circuit, (b) its Norton equivalent, and (c) the circuit consisting of the nonlinear current source.

$$V_{\mathrm{B}}(\omega_2 - \omega_1) = (I_{\mathrm{IN}}(\omega_2 - \omega_1) - i_{\mathrm{NL2}}(\omega_2 - \omega_1)) \cdot Z_{\mathrm{BB}}(\omega_2 - \omega_1) , \quad (7.1)$$

in which the nonlinear current can be further written as

$$i_{\mathrm{NL2}}(\omega_2 - \omega_1) = K_2 \cdot V_{\mathrm{B}}(\omega_2) \cdot V_{\mathrm{B}}(-\omega_1) . \quad (7.2)$$

However, it is not necessary to calculate the value of the nonlinear current source, because it can be observed from the measurements by tuning the $I_{\mathrm{IN}}(\omega_2 - \omega_1)$ and monitoring the $V_{\mathrm{B}}(\omega_2 - \omega_1)$. Once $V_{\mathrm{B}}(\omega_2 - \omega_1)$ is forced to zero, the following requirement is fulfilled

$$I_{\mathrm{IN}}(\omega_2 - \omega_1) = i_{\mathrm{NL2}}(\omega_2 - \omega_1) . \quad (7.3)$$

Since the impedance is defined as the ratio between the node voltage and the current, the impedance seen by the nonlinear current source can be modified by I_{IN}, and written as

$$Z_{\mathrm{BBeff}} = \frac{V_{\mathrm{B}}(\omega_2 - \omega_1)}{i_{\mathrm{NL2}}(\omega_2 - \omega_1)} = \left(\frac{I_{\mathrm{IN}}(\omega_2 - \omega_1)}{i_{\mathrm{NL2}}(\omega_2 - \omega_1)} - 1\right) \cdot Z_{\mathrm{BB}}(\omega_2 - \omega_1). \quad (7.4)$$

Equation (7.4) is explained in more detail in Figure 7.6, where Z_{BB} represents the original node impedance of the base. Once I_{IN} is applied according to (7.3), no envelope voltage waveform is seen at the base, which means that Z_{BBeff} is virtually driven to zero at the envelope frequency. Next, $I_{\mathrm{IN}}(\omega_2–\omega_1)$ is adjusted by 180º, which means that I_{IN} and i_{NL2} are summed up with the same phase. By studying (7.4), it can be seen that the effective Z_{BB} is now twice the actual base impedance Z_{BB}. This is also demonstrated in Figure 7.6. By tuning the amplitude and phase of $I_{\mathrm{IN}}(\omega_2–\omega_1)$, all impedances can be generated.

Even though the circuit used in this study is greatly simplified, the same principle also holds to more complicated systems. Actually, it does not matter where the distortion at the envelope frequency is generated, whether it is produced just by input nonlinearities or also fed back from the output. From the base node point of view, the entire BJT amplifier can be considered a Norton equivalent, consisting only of a base-emitter impedance and base-emitter nonlinearity, as illustrated in Figure 7.5. In this way, the tuned Z_{BB} is optimal from the overall nonlinearity point of view, and not just from that of the input nonlinearity.

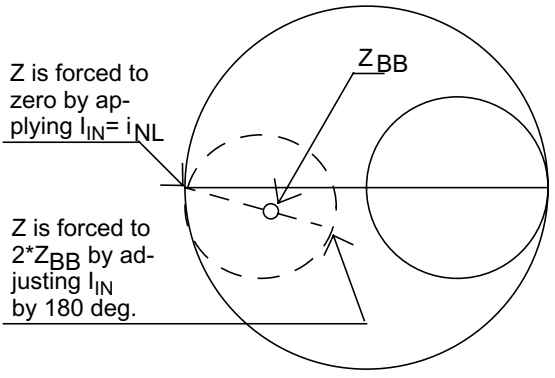


Figure 7.6 Z_{BB} generated by active load-pull. © IEEE 2000 [3].

7.2.2 Test Setup and Its Calibration

A test setup for optimizing Z_{BB} at the envelope frequency is presented in Figure 7.7. Three phase-locked signal generators are used, two for making the two-tone test and one low-frequency tone for modifying the value of Z_{BB} seen by the amplifier. The power of these tones is combined in a three-way power combiner and applied to the amplifier, and the output spectrum around the fundamental tones is monitored by a spectrum analyzer. The base signal at the envelope frequency is picked up without loading the base node by an adequate series resistor and bandpass filtering. The base signal needs to be monitored to find the condition of Z_{BB}=0 for calibration, which appears there when the envelope tone at the base disappears, as illustrated in Figure 7.6. The absolute values of the amplitudes of the signal generators in calibration and actual measurements are not important, because the impedance is calculated as a ratio of the two, as seen from (7.4). Instead, it is important to record the output signal as a voltage or current instead of the power to be used in calculations. Compared to most source-pull test setups, for example, [4], the one presented here is very simple and does not require any special measurement equipment.

From measurement accuracy point of view, the measured nominal value of Z_{BB} (without active loading) is important, because all the values are calculated based on this. Therefore, it has to be measured very carefully and since Z_{BB} varies according to the biasing conditions and the signal applied, it is important to use the same bias voltages in both Z_{BB} and the

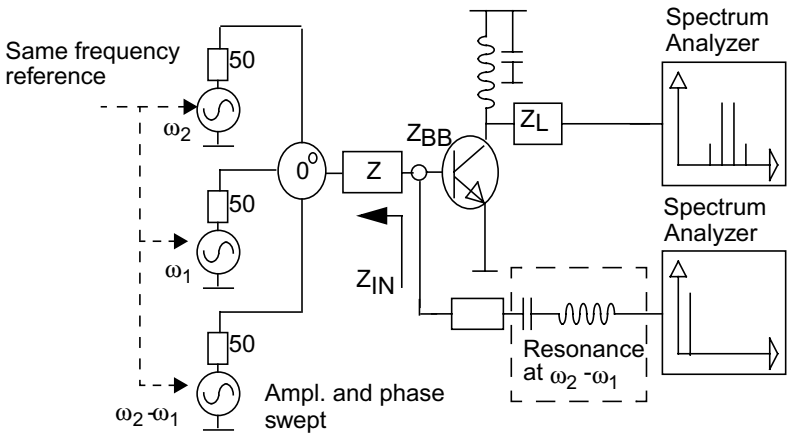


Figure 7.7 The test circuit for optimizing Z_{BB}(env). © IEEE 2000 [3].

source-pull measurements. Z_{BB} can be measured directly from the base using a network analyzer, or alternatively, the active and passive parts of it can be measured separately. In some cases it is advantageous to measure the input matching separately, and after that measure Z_{BB} through the input matching. Usually this is less prone to stability problems and measurement signal disturbance, but it cannot be applied accurately at very low frequencies if the input matching is ac-coupled. Therefore, a direct measurement is usually more accurate at very low frequencies, while the latter procedure is more accurate at higher frequencies.

Finally, one should note that these measurements do not necessarily reveal low-frequency stability problems. Due to the use of active source-pull, the IM2 distortion currents see a modified base impedance, but at other frequencies the amplifier still sees the original, nonmodified base impedance. Hence, the stability using the optimized bias impedance must be guaranteed by other means.

7.2.3 Optimum Z_{BB} at the Envelope Frequency Without Predistortion

The Philips BFG11 power BJT stage already used in many experiments in this book is applied at the center and modulation frequency of 1.8 GHz and 2 MHz to demonstrate the effects of input matching at the envelope frequency. As the matching of fundamental impedances is considered in detail in [5], the focus here is only on studying the out-of-band terminations. The power of the lower and higher IM3 signals in dBc at different Z_{BB}(env) values are presented in Figure 7.8(a). The figure shows that the IM3 values vary more than 15 dB, depending on Z_{BB}(env), and a significant linearity improvement from the original value of –33.5 dBc (marked with a circle) can be achieved by using the optimized Z_{BB}(env). The envelope impedance affects considerably the asymmetry of the IM3 sidebands, and the optimum input impedances for the two sidebands are different. There are two possible reasons why the optimum Z_{BB}(env) is different for upper and lower sidebands. First, the fundamental or second harmonic Z_{BB} may not be flat, and, second, the load terminations may not be optimal. IM3 improvements of up to 8 dB are obtained for both sidebands simultaneously, however, as shown in Figure 7.8(b), which shows the amplitude of the stronger IM3 component. This 8-dB improvement is very significant, because such a linearity improvement can be achieved by increasing the collector current by tens of percents, but this could be avoided simply by optimizing Z_{BB}.

The measurement system can also be employed for optimizing the impedance at the second harmonic as well, by tuning the series resonance and applying a signal source at that frequency. The effects of impedance at

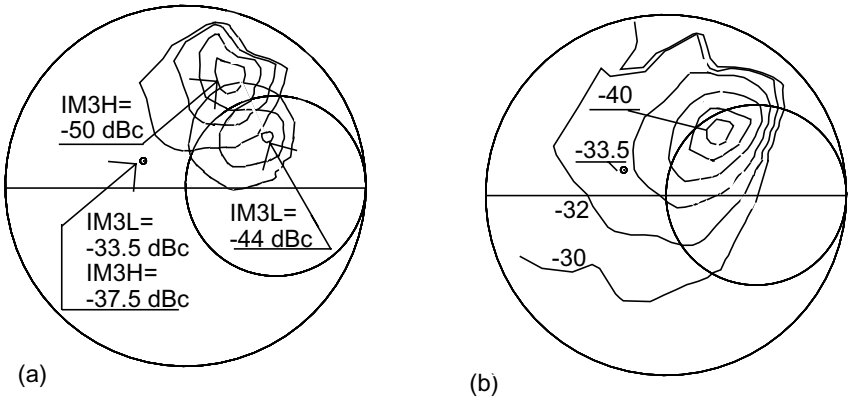


Figure 7.8 (a) Measured IM3L and IM3H at different Z_{BB}, and (b) a stronger IM3 at center and modulation frequencies of 1.8 GHz and 2 MHz. © IEEE 2000 [3].

the second harmonic can be summarized by stating that the second harmonic of the lower two-tone signal $2\omega_1$ affects mainly the IM3L signal and $2\omega_2$ the upper IM3H tone. This is evident, because nonlinearities of up to the third-degree behave in this way, as explained in Section 3.3. If higher degree nonlinearities play an important role, both second harmonics will mix with both IM3 frequencies. In most cases, however, minimization of IM3 by optimizing the input impedance inside the second harmonic band is not possible, because the center frequency changes from channel to channel. Since it varies considerably more than the maximum signal bandwidth, separate frequency bands for lower and higher second harmonics cannot be pinpointed.

7.2.4 Optimum Z_{BB} at the Envelope Frequency with Predistortion

This section discusses the optimum Z_{BB} with a polynomial RF predistorter. As in the previous section, the IM3 products are monitored while Z_{BB}(env) is virtually adjusted by the test setup in Figure 7.7.

Figure 7.9 shows the IM3L and IM3H products as a function of tone spacing in three different situations. The uppermost are the distortion products of the amplifier itself, without any linearization, while the middle curves present the IM3 using a memoryless RF predistorter, described in Appendix D. The predistorter is tuned at 1 MHz, and it works reasonably

well over a tone spacing from 1 to 7 MHz. IM3 stays below –49 dBc over this range, which corresponds to approximately 10-dB cancellation. Above 7 MHz, both of the sidebands increase rapidly and the predistorter loses its cancellation performance. At low modulation frequencies below 1 MHz, the IM3L increases rapidly while IM3H remains low. Altogether, canceled IM3 curves show that the amplifier exhibits very significant memory effects and wideband signals cannot be linearized successfully by means of a memoryless predistorter.

The bottom curve in Figure 7.9 presents the amplitude of the stronger canceled IM3 component at the output of the amplifier when Z_{BB}(env) is optimized separately at each tone spacing. Since the envelope impedance can be tuned to minimize the memory effects, the cancellation performance improves. The cancellation decrease caused by the memory effects at high and low tone spacings can be partially canceled out, and an IM3 level better than –49 dBc is achieved from dc up to 20 MHz by optimizing Z_{BB}(env). Without optimization, the same IM3 was achieved only from 1 to 7 MHz.

The original and optimized Z_{BB}(env) are depicted as functions of the modulation frequency in Figure 7.10. The original and optimum impedances are quite close to each other at low modulation frequencies, but in spite of that, IM3H can still be reduced by 10 dB using an optimum Z_{BB}(env). The original Z_{BB}(env) starts to decrease with increasing modulation frequency, while the opposite behavior is needed for maximum

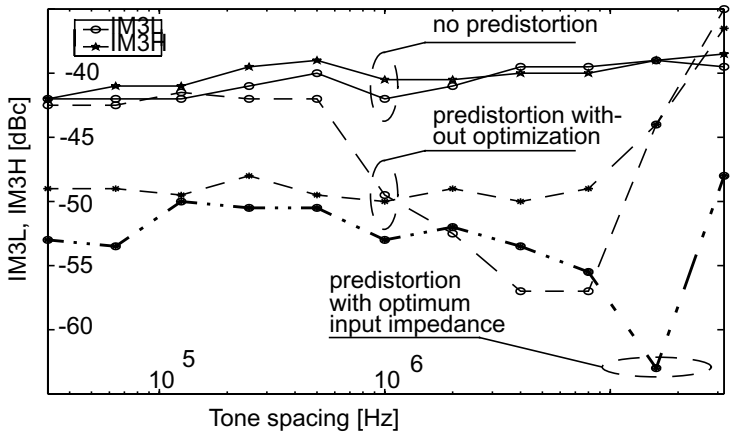


Figure 7.9 Measured amplitude of IM3 sidebands as a function of modulation frequency using a predistorter and optimized Z_{BB}. © IEEE 2000 [3].

linearity. If the envelope impedance at high modulation frequencies is optimized, the maximum correction bandwidth is increased from 7 to 20 MHz.

Then, how do we synthesize the optimal Z_{BB}(env) in practice? Z_{BB} is a parallel connection of the matching network impedance, bias circuitry, and the internal impedance of the transistor, which also is a function of the bias point. Since the internal impedance cannot be affected in most cases, Z_{BB} can be optimized by the input matching network, and especially if Z_{BB}(env) is to be optimized, by means of the input bias network.

Finally, it is important to emphasis that the dc bias impedances have stability constraints as well, as care is needed to guarantee the low-frequency stability. Moreover, the reader should be aware that the source-pull technique described above does not reveal stability problems, as from stability point of view the amplifier still sees the original impedances plus a single-tone source-pull signal in the input. It may be impossible to cancel the distortion completely by tailoring the bias impedance, but it may still be possible to flatten the memory effects by smaller changes in the impedances. Altogether, the bias impedances have a very significant effect to the linearity, memory effects, and stability properties of the amplifier, and they need to be designed very carefully to obtain the optimum performance.

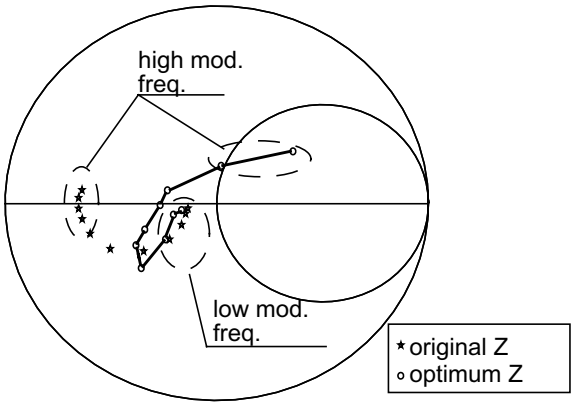


Figure 7.10 Measured original (stars) and optimized (solid) Z_{BB}(env) of the BJT amplifier. © IEEE 2000 [3].

7.3 Envelope Injection

The envelope injection technique is the third technique presented in this book to minimize memory effects, and it overcomes some of the drawbacks of the techniques presented earlier. The major drawbacks of the envelope filtering are tough accuracy requirements of the filtering and feasibility for symmetrical IM3 deviations only. Impedance optimization is also a good way to minimize memory effects, but its major problem is how to implement the optimized impedances in practice without endangering the stability of the amplifier.

An envelope frequency feedback method for linearizing an amplifier has been presented in [6], and some more recent solutions are presented in [7, 8]. In these papers a technique called the difference frequency technique is used for improving the IM3 performance of a PA without using any other linearization methods. The same principle of injecting out-of-band frequency components, but now at the second harmonic, is used in [9]. In this section, the envelope signal is injected to the input of the amplifier to minimize its memory effects, and the amplifier is further linearized using a memoryless RF predistortion. We call this the envelope injection technique.

One way of interpreting the envelope injection technique is to think of it as a real-time version of the source-pull technique presented in the previous section. Instead of applying an external envelope signal from a signal generator to the input of the amplifier, the envelope signal is generated here simply by squaring the input RF signal, hence producing the same spectra as quadratic input nonlinearities. This signal is then properly shaped and added to the input of the amplifier, where it actively modifies the impedance seen by the input second-order distortion currents.

The block diagram of the envelope injection technique with a polynomial third-order predistorter is given in Figure 7.11. As observed

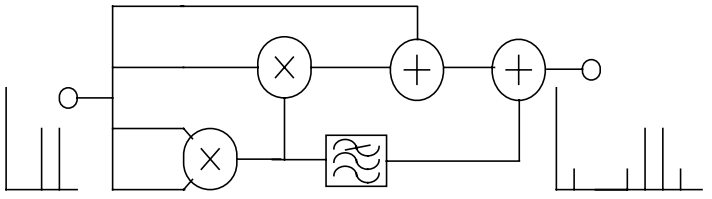


Figure 7.11 Principle of polynomial RF predistortion with envelope injection technique. © IEEE 2001 [1].

many times in this book, memory effects mostly arise as mixing products from the envelope frequency. In the envelope injection technique, only that part of the IM3 which is upconverting from the envelope frequency and causing memory effects is affected, and not the entire IM3 vector as in the envelope filtering technique. This difference is illustrated in Figure 7.12. In the envelope filtering technique, shown in Figure 7.12(b), the entire IM3 predistortion vector needs to be rotated to cancel the tiny memory effects, and this causes very strict accuracy requirements. In the envelope injection technique, the memoryless predistorter is tuned to cancel the memoryless part of the distortion, and the envelope injection is only used to minimize the frequency dependent part. This relaxes the accuracy requirements of the envelope filter.

Another important advantage of the envelope injection technique is that it is able to correct asymmetrical IM3 sidebands. It is commonly observed that the amplitudes of IM3L and IM3H are different and this phenomenon is mostly caused by the IM3 distortion upconverting from the envelope frequency. Since this part of the IM3 distortion can be corrected, or even zeroed by envelope injection, it is evident that this technique can be employed to improve the symmetry of IM3 sidebands.

Figure 7.13 presents the test setup for characterizing the frequency and amplitude response for the optimum envelope injection signal. It consists of a regular two-tone input with a third-order predistorter, a third signal generator for the injection signal, and a phase comparison block for phase calibration (see Appendix D). The procedure for characterizing the

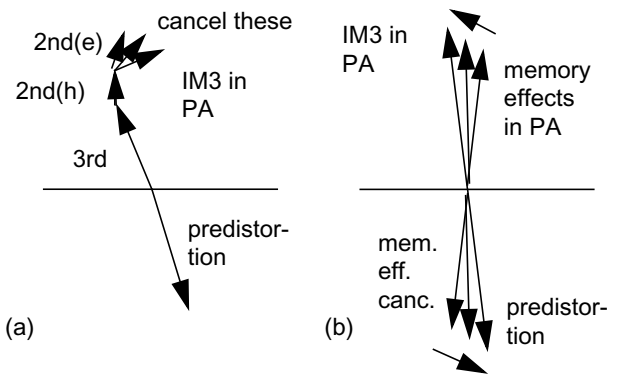


Figure 7.12 Compensation for memory effects using (a) the envelope injection technique, and (b) the envelope filtering technique. © IEEE 2001 [1].

optimum envelope injection signal is as follows: First, the predistorter is tuned without the injection signal at some modulation frequency. The selection of the frequency is quite critical, and should be chosen so that the memory effects are minimal at that frequency. Also, if the tuning of the predistorter is mismatched, the accuracy requirements for the injection signal will be increased. Second, the tone spacing is swept and the distortion products increase because of incomplete cancellation due to the memory effects. Now the injection signal is applied to nullify the increased distortion products, and the settings of the injection generator directly give the requirements for the injection signal. Once the modulation frequency and amplitude are swept, the requirements for the injection signal are obtained, and the injection signal can be replaced by the rectified envelope signal and a synthesized filter plus a possible nonlinear circuit for shaping the amplitude response, as discussed later.

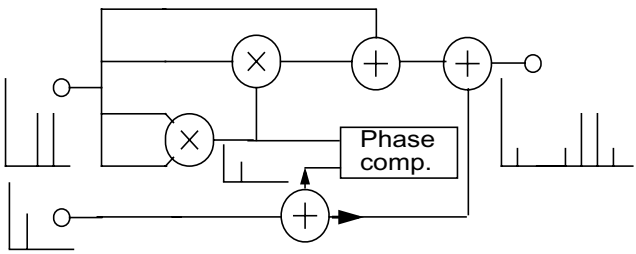


Figure 7.13 Test setup for characterizing the optimum envelope injection. © IEEE 2001 [10].

7.3.1 Cancellation of Memory Effects in a CE BJT Amplifier

Next, the envelope injection technique is studied using the same CE BJT amplifier as in the envelope filtering experiment. First, the required amplitude and phase for the envelope injection signal for a –50 dBc IM3 level are shown in Figure 7.14. These results are obtained by simulating the Volterra model of the BJT amplifier. To reach the linearity level of –50 dBc, the maximum allowable amplitude error is 5 dB and the maximum phase error 30º, as seen in Figure 7.14. This result indicates that the envelope injection is much less sensitive to filtering errors than the envelope filtering. This can be explained by looking again at Figure 7.12. Since only a small part of the IM3 distortion is affected, relative errors in that vector

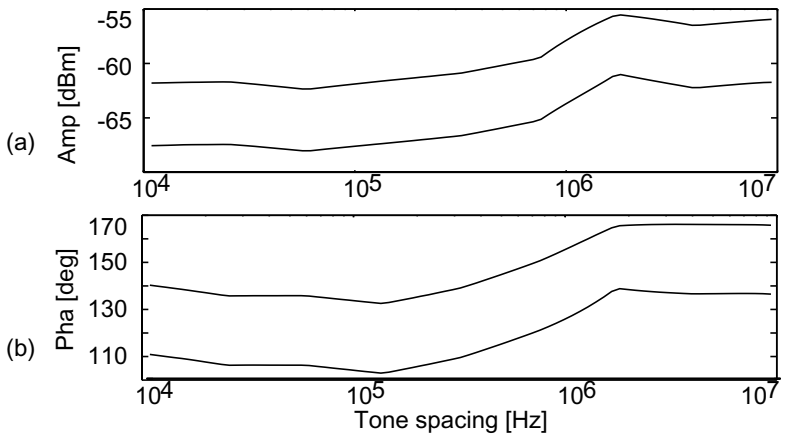


Figure 7.14 Volterra-simulated accuracy requirements for the (a) amplitude and (b) the phase of the envelope injection signal for –50 dBc IM3 levels. © IEEE 2001 [1].

have a weaker effect on the result. In the envelope filtering, however, the entire predistortion signal is adjusted, and since this is a large vector, even a small error in it will cause a significant error in the resultant vector.

Figure 7.15 presents the measured IM3 values of a BJT amplifier. The uppermost two curves are the IM3L and IM3H products of the amplifier alone, while the middle ones are the IM3 distortion products using a polynomial predistorter, tuned at 3 MHz modulation frequency. A 20-dB cancellation of IM3 is achieved at this point, but due to severe memory effects in the amplifier, the cancellation performance drops sharply if the tone spacing is either increased or decreased from 3 MHz. The lowest two curves are the IM3 products when the envelope injection signal is applied together with the polynomial predistortion, and the achieved cancellation performance stays good over the entire frequency range. The measured accuracy requirements of the injection signal for 20-dB cancellation are 3 dB and 20º, respectively, while the corresponding figures for envelope filtering were 0.5 dB and 3º. The measured accuracy differences in these two techniques are even larger than those expected on the basis of the simulations.

It is important to emphasize that the results shown in Figure 7.15 are carried out with an actual, implemented injection filter, not just with the signal source used as an injection signal, as in Figure 7.9. The implemented filter is as simple as a single series capacitor, causing a first-order highpass

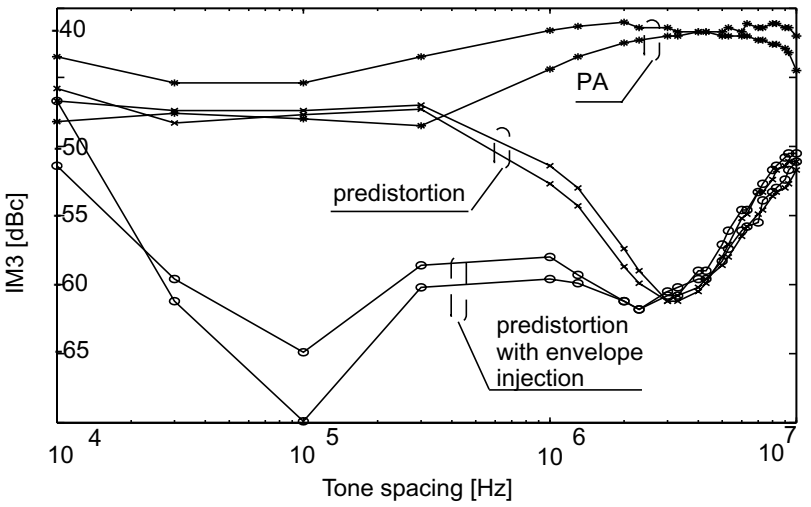


Figure 7.15 Measured improvement in IM3 performance in a BJT amplifier using no linearization, a memoryless predistorter, and predistorter with envelope injection. © IEEE 2001 [1].

response between the squaring circuit and the amplifier input. Since the filter is optimized below 3 MHz, no improvement above that frequency is achieved, but the filter may as well be designed for higher modulation frequencies, resulting in similar linearity improvements at higher modulation frequencies. Also, due to a highpass type of filter, the memory effects at very low modulation frequencies cannot be corrected by the presented filter.

7.3.2 Cancellation of Memory Effects in a CS MESFET Amplifier

Now the memory effects of the implemented CLY2 CS MESFET amplifier are minimized using the envelope injection technique. Compared to earlier simulations and measurements of the CLY2 amplifier, the resonance at 500 kHz is now removed by redesigning the drain bias network, but the shape of the high-frequency memory effects is still the same as before.

7.3.2.1 Frequency Domain Compensation

The third-order predistorter is tuned at center and modulation frequencies of 1.8 GHz and 320 kHz, respectively, given an output power level of 11 dBm. Compared with the BJT presented in the previous section, the

predistorter is now tuned at narrow tone spacing, resulting in a highpass type of injection filter and still retaining good cancellation down to very low modulation frequencies. This is possible because the MESFET does not have strong thermal memory effects. The IM3 components at the tuning point are reduced from –37 to –60 dBc, corresponding to a cancellation of 23 dB. The cancellation is given as a function of modulation frequency in Figure 7.16(a), and without envelope injection it decreases rapidly with increasing modulation frequency. This indicates that the amplifier exhibits strong high-frequency memory effects. Note that part of the memory effects seen in Figure 7.16 are caused by the predistorter, and for that reason the cancellation is markedly bandwidth-limited.

Much of the memory effects can be removed by envelope injection, however, and good linearity can be maintained up to higher bandwidths. In Figure 7.16(a) IM3 is always kept below –58 dBc by applying a proper envelope signal, and the required amplitude and phase windows of the injection signal are shown in Figure 7.16(b). The accuracy requirements are initially very loose, but tighten with increasing modulation frequency due to increasing IM level and hence higher cancellation required. The amplitude and phase windows show that a simple first-order highpass filter yields a properly shaped injection signal.

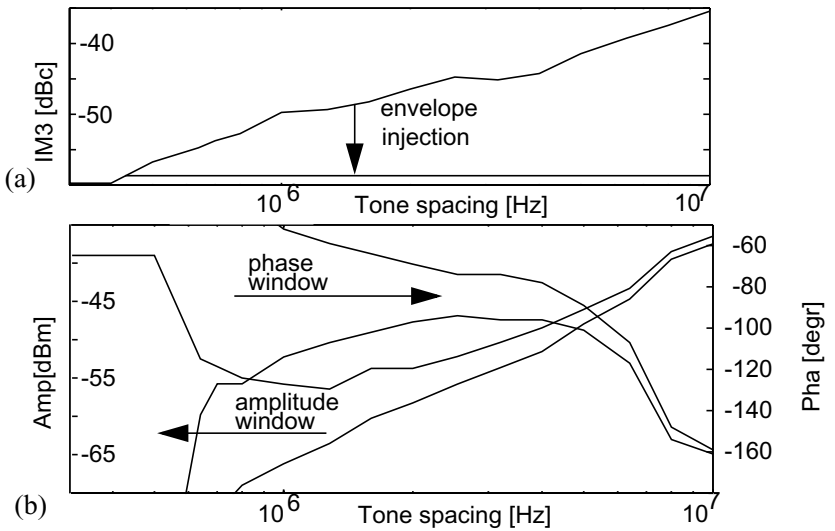


Figure 7.16 (a) Measured cancellation using a polynomial predistorter, and (b) amplitude and phase windows of the injected envelope signal for a 22-dB cancellation. © IEEE 2001 [10].

7.3.2.2 Amplitude Domain Compensation

As discussed in Chapters 3 and 6 the memory effects are not just frequency domain effects, but the signal amplitude also plays an important role, if higher than cubic nonlinearities are significant. This is the outcome of two facts: first, the frequency bands are wider at higher amplitudes, so that not just the second (ω_2-ω_1) but also a fourth-order envelope tone $2\omega_2$–$2\omega_1$ must be taken into account in a dc band, for example. Since the memory effects arise from nonflat frequency bands, wider bands emphasize these effects. Second, spectral components only up to the second harmonic convert down to the IM3 if the effects higher than third-order ones are negligible. However, if the fifth-order effects are significant, also the third harmonic will convert down to IM3. Both of these effects cause the frequency response of the memory effects to depend on the amplitude.

The compensation of amplitude domain memory effects is demonstrated here first by simulating the normalized IM3 components using the fifth-order model of the CLY2 MESFET amplifier. Owing to the redesigned drain bias circuit the 500 kHz resonance is now removed, and a new surface of the magnitude of the IM3L is shown in Figure 7.17(a). The amplitude sweep with a narrow tone spacing corresponds to the situation with low memory effects (as the thermal memory effects are not significant in the MESFET), and large deviations from this are observed at higher tone spacing. The envelope injection is applied in Figure 7.17(b), and most of the high-frequency memory effects caused by the amplifier can be compensated for.

Although the frequency domain effects can be minimized by the envelope injection technique, it is apparent from Figure 7.17(b) that the amount of distortion varies with the signal amplitude and cannot be completely canceled by a third-order predistorter. Better cancellation could be achieved using a fifth-order polynomial predistorter.

The measured results of an amplitude sweep from 12 to 15 dBm at a fixed 320 kHz tone spacing are shown in Figure 7.18(a). The solid and dashed curves represent the IM3 level with and without predistortion. The rapid linearity decrease caused by the fifth-order effects and memory effects is observed at high amplitude values, and the corresponding cancellation drops to as low as 5 dB at high amplitudes, as shown in Figure 7.18(b). Using the envelope injection, however, the cancellation can be maintained at 20 dB over the entire amplitude range, as indicated by the dashed line in Figure 7.18(b).

As the aimed 20-dB cancellation in Figure 7.18(b) is affected by the accuracy of the injected signal, next we will study the accuracy

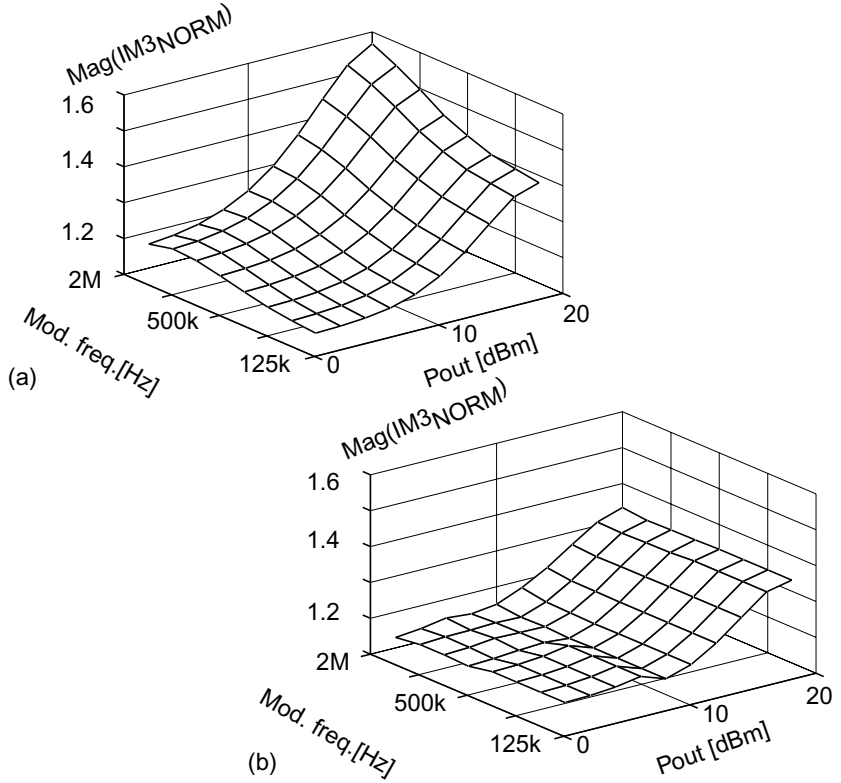


Figure 7.17 Simulated normalized magnitude of IM3L as a function of tone spacing and amplitude (a) without and (b) with envelope injection. © IEEE 2001 [10].

requirements. The solid lines in Figure 7.19(a) show the amplitude window of the injection signal, which gives a cancellation better than 20 dB over the amplitude range. Figure 7.19(b) shows the phase window for the same cancellation, showing tighter requirements with increasing memory effects. The requirements for a 10-dB cancellation improvement are no more than 2 dB and 20°, however, which are quite easy to attain.

The correct injection signal is easy to generate in the modulation frequency domain, simply by filtering the signal appropriately. In the amplitude domain however, the situation is slightly more complicated, because the downconverted injection signal is dependent on the input signal exactly to the power of two, and its phase is amplitude-independent. Figure 7.19(a), nevertheless, shows that an approximately 4:1 amplitude slope is

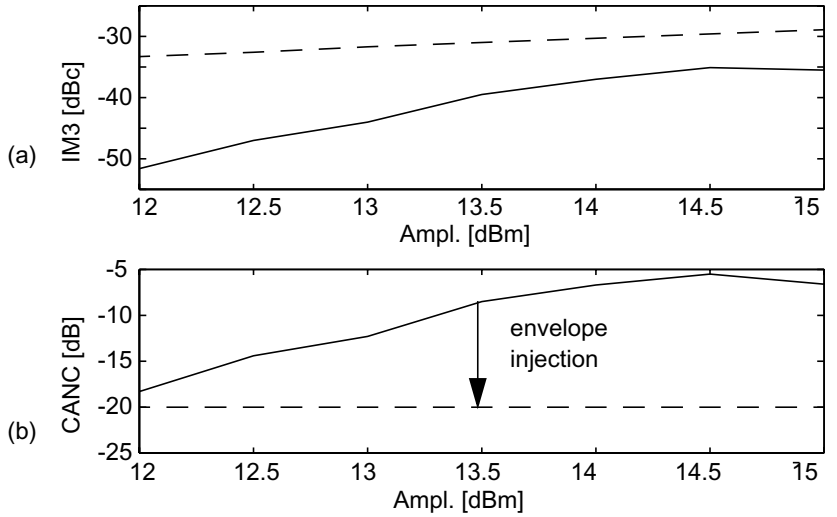


Figure 7.18 (a) Measured IM3 with (solid) and without (dashed) predistortion and (b) cancellation as functions of the fundamental output power, using a third-order predistorter with (dashed) and without envelope injection (solid). © IEEE 2001 [10].

required for the injection signal. This is difficult to achieve using analog components without distorting the signal and generating new spectral components. Although appropriate amplitude characteristics can be attained by digital signal processing (DSP), for example, we study here how a distorted injection signal including second- and fourth-order envelopes can be used to obtain the desired amplitude characteristics.

The circuit presented in Figure 7.20 is added to the injection path of the test setup presented in Figure 7.11, and measurements are carried out to characterize the injection signal. K4 is tuned so that the amplitude of the fourth-order injection signal is 6 dB lower than the fundamental injection at the maximum amplitude, and K2 is set at 1 to ensure that the measurements with and without the fourth-order envelope are comparable.

The measured reduction of memory effects achieved in these two cases are identical, but the requirements for the injection signals are different. The amplitude window for the injection signal is represented by the dashed line in Figure 7.19(a). The figure also shows the amplitude required for the fundamental injection signal. The amplitude of the fourth-order signal can be seen to vary with amplitude, and is 6 dBc at the maximum level used in our experiments. The amplitude slope required for the fundamental

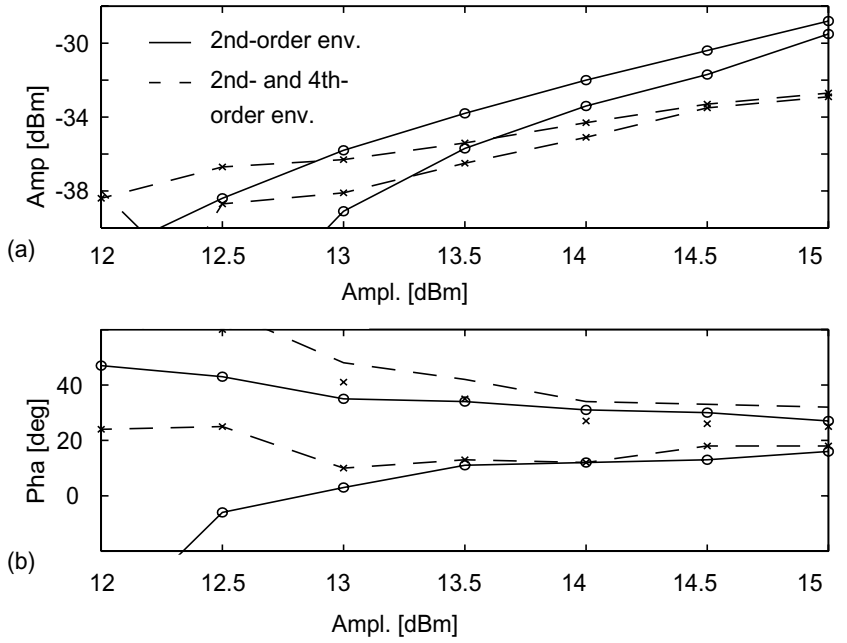


Figure 7.19 (a) Measured amplitude and (b) phase requirements for envelope signal using second-order (solid) and both second and fourth-order (dashed) envelope injection. © IEEE 2001 [10].

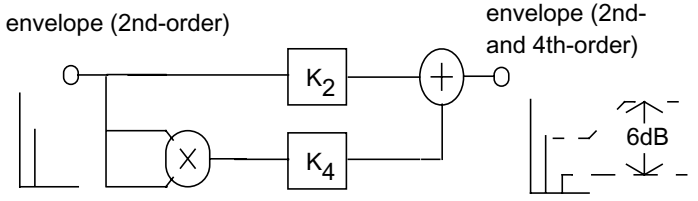


Figure 7.20 Block diagram of the circuit producing the fourth-order envelope. © IEEE 2001 [10].

injection is very close to the optimum at 2:1, making it easy to implement it. There is a drawback, however, since the amplitude window is narrower than that for the injection without the fourth-order signal.

7.4 Summary

Memory effects in power amplifiers reduce the cancellation performance of predistortion linearization. Three techniques to minimize the memory effects, envelope filtering, impedance optimization, and envelope injection, are presented in this chapter, and improvements in cancellation performance are achieved with all of the techniques, indicating that the memory effects can be removed at least partially. There are significant differences in how easily the techniques can be implemented, however.

In the envelope filtering technique, inverse memory effects compared to the ones in the amplifier are generated inside the predistortion device. This is done by filtering and phase-shifting the rectified envelope signal, and the technique can be used with polynomial or complex gain predistortion, where the envelope signal is used to generate the IM3 sidebands.

Impedance optimization and envelope injection both attack the baseband bias impedances seen by the distortion current sources. Most of the memory effects are caused by the upconversion of the rectified second- (and fourth-) order envelope signal, and the frequency response and amplitude of this upconverted envelope term can be affected by optimizing or actively controlling the low-frequency input impedance Z_{BB} or Z_{GG}. Minimization of the IM3 caused by the rectified envelope also improves the symmetry of IM3 sidebands, in which case a normal memoryless predistortion works more efficiently.

Impedance optimization is then based simply on the optimization of the out-of-band impedances. By optimizing the input impedance at the envelope frequency (from dc to 10 MHz to 20 MHz), most of the memory effects can be minimized. To find the optimum bias impedance, a source-pull test setup is developed, but practical implementation of this optimal impedance may be more difficult. The importance of out-of-band input terminations without a predistorter was also demonstrated by the measurements, and significant differences in optimum input impedance at the envelope frequency with and without predistortion were observed. In other words, PAs for standalone and linearized configurations have to be designed differently. In a standalone operation, the optimization of the input impedance at the envelope frequency is determined simply by the minimization of IM3, whereas with predistortion the main target is to

minimize the memory effects and maintain symmetrical IM3 sidebands that can be further canceled using predistortion.

In the envelope injection technique, a low-frequency envelope signal is generated and added to the RF carrier, or more elegantly, the envelope signal modulates the input dc bias voltage. This signal is shaped so that the apparent input impedance seen by the input distortion current generators is either small enough or flat enough. The optimum frequency response for the envelope signal is obtained by using an external envelope signal, the phase and amplitude of which is varied to minimize the memory effects. The accuracy requirements for an injection filter are quite loose compared with those for other techniques, and often, a simple highpass stage implemented with a single series capacitance is sufficient. Here, envelope injection is used with polynomial RF predistortion, but it can be used with any kind of linearization technique.

The envelope injection technique can be used to reduce both modulation frequency and amplitude domain memory effects. In the frequency domain, a filter is synthesized, based on the measured optimum injection signal. The required accuracy of the filter depends on the amount of improvement desired, but in general quite large deviations in the response are tolerated. In the amplitude domain, cancellation is limited by higher order effects and memory effects. Since only a third-order predistorter was used but the amplifier shows a significant amount of distortion due to fifth-degree nonlinearities, the requirements for the injection to keep the cancellation good over the range of amplitudes are stringent. It can be expected, however, that the requirements will become looser when a fifth-order predistorter is used with envelope injection. In this way a second-order envelope will mostly be used to compensate for the memory of the IM3 products, while a fourth-order envelope will be used mostly for the IM5 terms. Since higher order products affect lower order responses, optimization has to be performed in decreasing degree of nonlinearity (i.e., the highest order products are first compensated).

The envelope signal is generated in this chapter as a mixing product, the alternative being to take the properly shaped modulated data directly from the DSP. Digital filtering is preferable over analog filtering, because the latter may be inaccurate and requires a large board or chip area. DSP would also give numerous degrees of freedom in signal processing, because more complex power dependence relations can be realized without distorting the injection signal. This is particularly important in amplitude domain compensation.

7.5 Key Points to Remember

1. The cancellation performance of predistortion is sensitive to memory effects generated in RF power amplifiers.
2. Predistortion works best, if the IM3 sidebands are symmetrical, that is, the amplitudes of IM3 tones are the same, and if their phase depends on the difference to the center frequency, the upper and lower sidebands see opposite phase shifts.
3. Memory effects can be minimized using envelope filtering, impedance optimization, and envelope injection techniques.
4. In the envelope filtering technique, opposite memory effects are built inside the predistortion device.
5. Since the electrical memory effects are caused by varying impedances inside the frequency bands, their optimization will reduce the memory effects.
6. Optimal terminal impedances of the amplifier differ in standalone and linearized configurations: the absolute value of IM3 is the main interest without linearization, while the amount of memory effects is the most important thing with linearization.
7. In the envelope injection technique, a properly shaped envelope signal is added to the input of the amplifier.
8. Since most of the memory effects are mixed from the envelope frequency, only the part of the distortion that exhibits memory effects is compensated in the envelope injection technique, and therefore the accuracy requirements for the injection signal are quite loose.
9. The envelope filtering cannot correct the asymmetry between the IM3 sidebands. Instead, impedance optimization and the envelope injection technique can improve the symmetry of the distortion as well.

References

[1] Vuolevi, J., J. Manninen, and T. Rahkonen, "Cancelling the memory effects in RF power amplifiers," *Proc. of IEEE International Symposium of Circuit and Systems* (*ISCAS01*), Sydney, Australia, May 6-9, 2001, Proceedings Vol. I, pp. I.57-I.60

[2] Cripps, S., *RF Power Amplifiers for Wireless Communications*, Norwood, MA: Artech House, 1999.

[3] Vuolevi, J., T. Rahkonen, and J. Manninen, "Measurement technique for improving linearity by optimizing the source impedance of RF power amplifiers," *Proc. 2000 IEEE Radio and Wireless Conference (RAWCON00)*, Denver, CO, September 10-13, 2000, pp. 227-230.

[4] Berghoff, G., et al., "Automated characterization of HF power transistors by source-pull and multiharmonic load-pull measurements based on six-port techniques," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 46, No. 12, 1998, pp. 2068-2073.

[5] Iwai, T., et al., "High efficiency and high linearity InGaP/GaAs HBT power amplifiers: matching techniques of source and load impedance to improve phase distortion and linearity," *IEEE Trans. on Electron Devices*, Vol. 45, No. 6, 1998, pp. 1196-1200.

[6] Hu,Y., J. Mollier, and J. Obregon, "A new method of third-order intermodulation reduction in nonlinear microwave systems," *IEEE Trans. on Microwave Theory and Techniques*, Vol. 34, No. 2, 1986, pp. 245-250.

[7] Modeste, M., et al., "Analysis and practical performance of a difference frequency technique for improving the multicarrier IMD performance of RF amplifiers," *Proc. 1999 IEEE MTT-S Symposium on Technologies for Wireless Applications*, pp. 53-56.

[8] Jenkins, W., and A. Khanifar, "A multicarrier amplifier with low third-order intermodulation distortion," *2000 IEEE MTT-S International Microwave Symposium Digest*, Vol. 3, pp. 1515-1518.

[9] Joshin, K., et al., "Harmonic feedback circuit effects on intermodulation products and adjacent channel leakage power in HBT power amplifier for 1.95 GHz wide-band CDMA cellular phones," *IEICE Transactions on Electron*, Vol. 82, No. 5, 1999, pp. 725-729.

[10] Vuolevi, J., J. Manninen, and T. Rahkonen, "Memory effects compensation in RF power amplifiers using envelope injection technique," *Proc. 2001 IEEE Radio and Wireless Conference (RAWCON01)*, Denver, CO, August 2001, pp. 257-260.

Appendix A: Basics of Volterra Analysis

This appendix discusses the Volterra series analysis in some more detail, mostly based on [1]. The output of a nonlinear system, with certain restrictions can be expressed with the following equation

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} h_1(\tau)x(t-\tau)d\tau_1 \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_2(\tau_1, \tau_2)x(t-\tau_1)x(t-\tau_2)d\tau_1 d\tau_2 \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_3(\tau_1, \tau_2, \tau_3)x(t-\tau_1)x(t-\tau_2)x(t-\tau_3)d\tau_1 d\tau_2 d\tau_3 \\ &+ \ldots + x(t-\tau_n)d\tau_1 d\tau_2 \ldots d\tau_n + \ldots \\ &= H_1[x(t)] + H_2[x(t)] + H_3[x(t)] + \ldots + H_n[x(t)] + \ldots \end{aligned} \tag{A.1}$$

The first term in the series is recognized as the normal convolution integral, describing the linear response of a system with memory. The following terms stand for nonlinear effects. This series is called the Volterra series and the n-dimensional impulse responses $h_n(\tau_1, \tau_2)$ are called the *Volterra kernels* of the system, and the $H_n[x(t)]$ are called nth-order Volterra operators. In this approach, the nonlinear system is considered as a combination of operators of different order. To demonstrate the use of the Volterra series approach for nonlinear calculations, let us assume that the input signal of a system is a single-tone sinewave that can be rewritten as a sum of two phasors $x_a(t)$ and $x_b(t)$

$$x(t) = A \cdot \cos(\omega_1 \cdot t) = \frac{A}{2} \cdot e^{j\omega_1 t} + \frac{A}{2} \cdot e^{-j\omega_1 t} = x_a(t) + x_b(t) \tag{A.2}$$

A linear system can be calculated using phasors and the response of the second-order system to this input is given by

$$\begin{aligned} y(t) &= H_2[x_a(t),x_a(t)] + H_2[x_b(t),x_b(t)] \\ &+ H_2[x_a(t), x_b(t)] + H_2[x_b(t), x_a(t)] \end{aligned} \tag{A.3}$$

where all possible combinations of picking up two signals to the second-order operator are listed; in general, n input signals are always needed for an nth-order operator. The first term in (A.3) can now be written in terms of the second-order Volterra kernel using the two-dimensional convolution

$$\begin{aligned} H_2[x_a(t),x_a(t)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_2(\tau_1, \tau_2)x_a(t-\tau_1)x_a(t-\tau_2)d\tau_1 d\tau_2 \\ &= \frac{A^2}{4}\cdot\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_2(\tau_1, \tau_2)e^{j\omega_1(t-\tau_1)}e^{j\omega_1(t-\tau_2)}d\tau_1 d\tau_2 \\ &= \frac{A^2}{4}\cdot e^{j2\omega_1 t}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h_2(\tau_1, \tau_2)e^{j\omega_1(-\tau_1)}e^{j\omega_1(-\tau_2)}d\tau_1 d\tau_2 \\ &= \frac{A^2}{4}\cdot H_2(j\omega_1, j\omega_1)\cdot e^{j2\omega_1 t} \end{aligned} \tag{A.4}$$

where $H_2(j\omega_1,j\omega_2)$ is the two-dimensional Fourier transform of the impulse response $h_2(t_1,t_2)$.

The second term of (A.3) can be calculated similarly by

$$H_2[x_b(t),x_b(t)] = \frac{A^2}{4}\cdot H_2(-j\omega_1, -j\omega_1)\cdot e^{-j2\omega_1 t}. \tag{A.5}$$

The third and fourth terms of (A.3) are identical in symmetrical systems and can be expressed by

$$H_2[x_a(t),x_b(t)] = \frac{A^2}{4}\cdot H_2(j\omega_1, -j\omega_1) \tag{A.6}$$

and

$$H_2[x_b(t),x_a(t)] = \frac{A^2}{4}\cdot H_2(-j\omega_1, j\omega_1) \tag{A.7}$$

The first and last two terms are complex conjugates to each other, and the output of the second-order Volterra kernel becomes

$$y_2(t) = \frac{A^2}{2} \cdot Re(H_2(j\omega_1, j\omega_1) \cdot e^{j2\omega_1 t}) \quad . \tag{A.8}$$
$$+ \frac{A^2}{2} \cdot Re(H_2(j\omega_1, -j\omega_1) \cdot e^{j0t})$$

Thus, the second-order response of $x(t)$ is written in a familiar manner as a product of the two-dimensional frequency response $H_2(j\omega_1,j\omega_2)$ and the distorted spectrum of the input signal, now consisting only of the second harmonic and the dc component. As time domain squaring (distortion) corresponds to frequency domain convolution, the distorted signal spectrum that will be multiplied by $H_2(j\omega_1,j\omega_2)$ is obtained simply by convolving (in the frequency domain) the two-sided spectrum $X(j\omega)$ of the input signal x(t) once with itself. In the case of a single-tone signal this generates the dc and second harmonic components, as seen above.

Then what does the two-dimensional frequency response function look like? As a very simple example, consider the cascade of a filter, quadratic nonlinearity and another filter shown in Figure A.1(a). For the second-order response, two inputs at frequencies ω_1 and ω_2 (that may be the same) are required, and the overall response of this circuit is shown in (A.9).

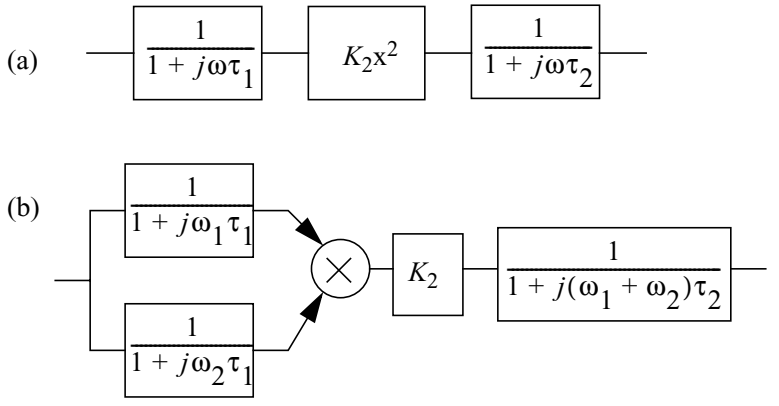


Figure A.1 (a) A cascade of input and output filters and a second-degree nonlinearity, and (b) the corresponding block diagram.

$$H_2(j\omega_1, j\omega_2) = \frac{K_2}{(1 + j\omega_1\tau_1) \cdot (1 + j\omega_2\tau_1) \cdot (1 + j(\omega_1 + \omega_2)\tau_2)} \quad \text{(A.9)}$$

Equation (A.9) makes sense in a two-tone case, too: Both input tones ω_1 and ω_2 are filtered with time constant τ_1 before entering the nonlinearity. Here, a distortion product at frequency $\omega_1 + \omega_2$ is generated, and it is further filtered with time constant τ_2 in the output filter. Thus, (A.9) includes the frequency response both before and after the nonlinearity. Note that the use of positive frequencies alone results in tones in the second harmonic band only, and also negative-frequency phasors $-\omega_1$ or $-\omega_2$ are needed to get the IM distortion products below the second harmonic band.

The simple introduction above illustrates many things. First, distortion clearly generates new frequency components, and the output spectrum of an *n*th-degree nonlinearity is simply an *n*-fold convolution of the input spectrum. The fundamental idea of polynomial modeling is that the spectrum of the distortion generated by each degree of nonlinearity can be quite easily calculated. In the case of a few discrete tones, we can relatively easily keep track of all the mechanisms that result in distortion on a particular frequency. This is a unique property of the Volterra analysis, and the main reason for using it in this book.

Second, the Volterra transfer functions shown above may be handy for modeling reasonably simple input-output nonlinearities, but in the case of multiple nonlinearities and feedback loops the transfer functions may well turn out to be complicated. Fortunately, in circuit analysis we do not need to derive the Volterra kernels by hand, but we can use the nonlinear current method (called the direct method in [1]) instead, which is conceptually very similar to noise analysis: we simply add distortion current sources in parallel with the nonlinear elements and calculate their response to the output. This makes it possible to build a per-component (termwise) plot of the distortion, exactly as we like to see dominant noise sources. The only big difference compared to noise analysis is that now all signals are correlated, and can cancel each other. Hence, the magnitude presentation used for finding dominant noise contributions is not sufficient, but also the relative phase information of the distortion contributions is necessarily needed. Luckily, phasor calculations automatically contain the phase information.

Third, the analysis is conceptually quite simple. We calculate (filter) the excitations of the nonlinearities, see what distortion tones are produced, and then filter them again on their way to the output. The only complication arises from the fact that the higher order products can be generated in tens

of different ways, and we need to calculate them all. In this book, the different mechanisms resulting in IM3 in a single-transistor common emitter stage have been constructed by hand and listed in Appendix C. It is, however, possible to obtain almost the same resolution of distortion analysis entirely numerically, and in that case we can analyze more complicated circuits and multitone signals as well.

Fourth, series expansions must always be truncated at some level. The truncation effects are studied by some examples in Appendix B, and the distortion analysis in this book is mostly limited to the third order, and occasionally increased to the fifth order to see some amplitude dependent phenomena. The reader should be aware that a third-order analysis predicts IM3 that never saturates but steadily increases in proportion to the third power of the input amplitude. In this case, we can never see IM3 dropping at a certain amplitude level, as this effect is due to the fifth- or higher order distortion that locally cancels the third-order distortion. If we want to catch this phenomenon, or to see how the relative phase of IM3 varies with signal amplitude, for example, a higher order model must necessarily be used.

Reference

[1] Wambacq, P., and W. Sansen, *Distortion Analysis of Analog Integrated Circuits*, Norwell, MA: Kluwer, 1998.

Appendix B: Truncation Error

In general, the nonlinear function can be presented by its Taylor series expansion. This polynomial consists of an infinite number of terms, however, and since in all practical situations the polynomial must be truncated, some amount of truncation error always exists. This figure is dependent on the amount of nonlinearity, the amplitude range to be covered, and the number of terms to be taken into account.

We will look at the truncation error using two common nonlinear functions. The first is the purely exponential collector current equation of a BJT. The second is the drain current equation of the FET, exhibiting mostly second-degree nonlinearity. The BJT collector current can be written as

$$I_C = I_S \cdot e^{\left(\frac{V_{BE} + v_{be}}{V_t}\right)}, \tag{B.1}$$

the up to fifth-degree Taylor expansion of which can be written by

$$I_C = I_S \cdot e^{(V_{BE}/V_t)} \cdot \left(1 + \frac{v_{be}}{V_t} + \frac{1}{2} \cdot \left(\frac{v_{be}}{V_t}\right)^2 + \frac{1}{6} \cdot \left(\frac{v_{be}}{V_t}\right)^3 + \frac{1}{24} \cdot \left(\frac{v_{be}}{V_t}\right)^4 + \frac{1}{120} \cdot \left(\frac{v_{be}}{V_t}\right)^5 + \ldots\right) \tag{B.2}$$

Figure B.1(a) presents the ac values of the collector current as a function of ac base voltage. Since the dc base voltage affects only the magnitude of the series, the nonlinearity is independent of the dc value of V_{BE}. Three curves are drawn in Figure B.1(a): the actual exponential function and two Taylor series of it, of which the first is truncated to the degree of three and the second the degree of five. Figure B.1(b) presents the relative errors of the polynomials compared to the actual function. The third-degree polynomial yields the maximum error of 5.5% at the normalized signal amplitude

v_{be}/V_t of 1, while the respective number for a fifth-degree one is 0.2%. If the amplitude is increased to 1.5, the fifth-degree polynomial yields the error of less than 1.7%.

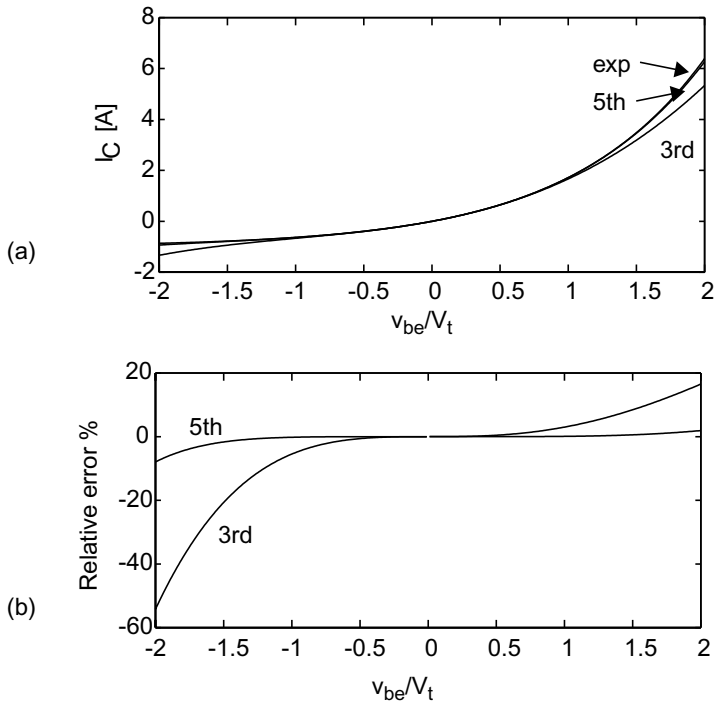


Figure B.1 (a) The actual exponential curve and its truncated polynomials of degrees three and five, and (b) relative error caused by truncation to the third and fifth degrees.

Next, the errors of the Taylor series expansion of the drain current equation are calculated. In a saturation regime the drain current can be expressed as

$$I_{DS} = \frac{\mu_0}{1 + \theta(V_{GSQ} + v_{gs} - V_T)} C'_{OX} \frac{W}{L}(V_{GSQ} + v_{gs} - V_T)^2 \qquad \text{(B.3)}$$

where μ_0, θ, and V_T are 0.047A/V^2, 0.0791/V, and 0.75V in this example. These coefficients affect the amount of nonlinearity, and moreover, the velocity saturation term in the denominator is now responsible for also generating other than plain second-degree nonlinearity.

The Taylor series expansion of (B.3) now depends on the dc value of V_{GS}, chosen to be 2V in this example. The nonlinearity coefficients (normalized by the linear term a_1) become a_2'=a_2 /a_1=0.347, a_3'= –0.025, a_4'=0.0018, a_5'= –0.00013. Figure B.2(a) presents the ac value of the drain current as a function of gate voltage. The actual function and its third- and fifth-degree expansions exist in Figure B.2(a), while Figure B.2(b) presents the relative errors. The third-degree polynomial yields the maximum error of 0.31% at the signal amplitude of v_{gs}=2Vpp. At the full signal level of 4 Vpp, the maximum errors of third- and fifth-degree models are 8.9% and 0.18%, respectively.

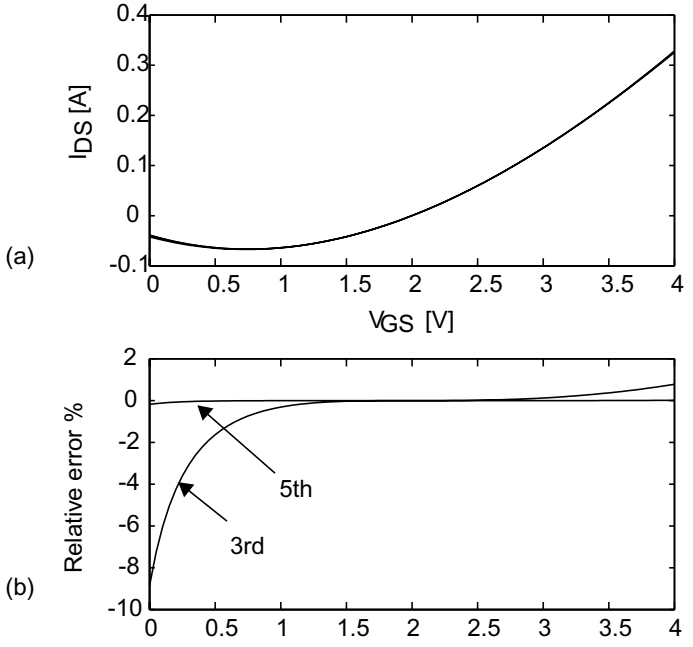


Figure B.2 (a) The drain current curve and its best fitted polynomials of degrees three and five, and (b) relative error caused by truncation to the third and fifth degree.

The above calculations show that the truncation errors are not that bad. The fifth-degree polynomial is quite accurate over a large amplitude range, and even the third-degree one can be successfully applied at reasonably high amplitude levels, especially in FET amplifiers in which the nonlinearities are not very strong. Actually, the original function to be fitted is not accurate either at low gate voltage values which exhibits the biggest deviations between the original and polynomially modeled drain current.

As a conclusion, the errors in a drain current are relatively small and even a strongly nonlinear collector current can be modeled reasonably accurately up to v_{be}/V_t ratios of higher than one (i.e., v_{be} larger than 26 mV).

Appendix C: IM3 Equations for Cascaded Second-Degree Nonlinearities

The equations for IM3 caused by cascaded second-degree nonlinearities become quite complicated, and some formalism has to be developed, and to combine complicated terms that will appear several times in the equations, certain transfer functions are derived. First, the term

$$TF(s) = \frac{v_{\mathrm{CE}}(s)}{v_{\mathrm{BE}}(s)} = \frac{g_m(Y_E + Y_L) + Y_{BE} \cdot Y_L - Y_E \cdot Y_{BC}}{Y_E \cdot (Y_L + Y_{CE} + Y_{BC}) + Y_{CE} \cdot Y_L}. \tag{C.1}$$

is handy for replacing v_{CE} with TFv_{BE} (note that all admittances Y may be functions of frequency as well). To describe the conversion from distortion currents to node voltages, transimpedance transfer functions are handy, and here, TF_{XYZ} means the transfer function from a nonlinear current between nodes X and Y to the voltage in node Z. For example, the second-order voltages at the base, collector, and emitter node can be calculated as

$$\begin{aligned} &V_{B2}(s_1 + s_2) \\ &= TF_{\mathrm{BEB}}(s_1 + s_2) \cdot (i_{NL2GPI} + i_{NL2CPI} + i_{NL2GPT} + i_{NL2CPT}) \\ &+ TF_{\mathrm{CEB}}(s_1 + s_2) \cdot (i_{NL2GM} + i_{NL2GMT}) \\ &+ TF_{\mathrm{CBB}}(s_1 + s_2) \cdot (i_{NL2CBC} + i_{NL2CBCT}) \end{aligned} \tag{C.2}$$

$$\begin{aligned} &V_{C2}(s_1 + s_2) \\ &= TF_{\mathrm{BEC}}(s_1 + s_2) \cdot (i_{NL2GPI} + i_{NL2CPI} + i_{NL2GPT} + i_{NL2CPT}) \\ &+ TF_{\mathrm{CEC}}(s_1 + s_2) \cdot (i_{NL2GM} + i_{NL2GMT}) \\ &+ TF_{\mathrm{CBC}}(s_1 + s_2) \cdot (i_{NL2CBC} + i_{NL2CBCT}) \end{aligned} \tag{C.3}$$

and

$$
\begin{aligned}
&V_{E2}(s_1+s_2) \\
&= TF_{BEE}(s_1+s_2)\cdot(i_{NL2GPI}+i_{NL2CPI}+i_{NL2GPT}+i_{NL2CPT}) \\
&+ TF_{CEE}(s_1+s_2)\cdot(i_{NL2GM}+i_{NL2GMT}) \\
&+ TF_{CBE}(s_1+s_2)\cdot(i_{NL2CBC}+i_{NL2CBCT})
\end{aligned}
\tag{C.4}
$$

where the distortion currents have been separated into purely electrical and electrothermal parts. Altogether, nine such transfer functions are needed, and they are listed below:

$$
\begin{aligned}
TF_{BEB}(s) &= \frac{V_B}{i_{BE}} \\
&= \frac{-[Y_E\cdot Y_L + Y_{CE}\cdot Y_E + Y_{CE}\cdot Y_L + gm\cdot Y_L + s\cdot C_{BC}\cdot Y_E]}{det(s)}
\end{aligned}
\tag{C.5}
$$

$$
\begin{aligned}
TF_{BEE}(s) &= \frac{V_E}{i_{BE}} \\
&= \frac{[Y_{IN}\cdot Y_L + Y_{CE}\cdot Y_{IN} - gm\cdot Y_L + s\cdot C_{BC}\cdot (Y_L + Y_{IN})]}{det(s)}
\end{aligned}
\tag{C.6}
$$

$$
\begin{aligned}
TF_{BEC}(s) &= \frac{V_C}{i_{BE}} \\
&= \frac{[Y_{CE}\cdot Y_{IN} + gm\cdot Y_{IN} + gm\cdot Y_E - s\cdot C_{BC}\cdot Y_E]}{det(s)}
\end{aligned}
\tag{C.7}
$$

$$
TF_{CEB}(s) = \frac{V_B}{i_{CE}} = \frac{[Y_{BE}\cdot Y_L + s\cdot C_{BC}\cdot Y_E]}{det(s)} \tag{C.8}
$$

$$
\begin{aligned}
TF_{CEE}(s) &= \frac{V_E}{i_{CE}} \\
&= \frac{[Y_{IN}\cdot Y_L + Y_{BE}\cdot Y_L + s\cdot C_{BC}\cdot (Y_{IN} + Y_L)]}{det(s)}
\end{aligned}
\tag{C.9}
$$

$$TF_{CEC}(s) = \frac{V_C}{i_{CE}} = \frac{-[Y_{IN}(s) \cdot Y_E(s) + Y_{BE} \cdot Y_{IN} + Y_{BE} \cdot Y_E + s \cdot C_{BC} \cdot Y_E]}{det(s)} \tag{C.10}$$

$$TF_{CBB}(s) = TF_{CEB}(s) - TF_{BEB}(s) \tag{C.11}$$

$$TF_{CBE}(s) = TF_{CEE}(s) - TF_{BEE}(s) \tag{C.12}$$

$$TF_{CBC}(s) = TF_{CEC}(s) - TF_{BEC}(s) \tag{C.13}$$

where

$$\begin{aligned} det(s) = {} & [Y_{BE} \cdot Y_{CE} \cdot (Y_L + Y_E + Y_{IN}) + Y_{IN} \cdot Y_L \\ & \cdot (Y_{BE} + Y_{CE} + g_m + Y_E) + Y_{CE} \cdot Y_E \cdot Y_{IN} + Y_{BE} \cdot Y_L \cdot Y_E \\ & + Y_{BC} \cdot [Y_{BE} \cdot Y_{IN} + Y_{CE} \cdot Y_{IN} + Y_E \cdot Y_{IN} + Y_{BE} \cdot Y_E \\ & + Y_{BE} \cdot Y_L + Y_{CE} \cdot Y_L + Y_E \cdot Y_L + Y_{CE} \cdot Y_E + gm \cdot Y_L \\ & + g_m Y_{IN} + g_m \cdot Y_E]] \end{aligned} \tag{C.14}$$

Further, from the transfer functions above we can derive transfer functions TF_{XYZW} that describe how the current between nodes X and Y translates to the voltage between nodes Z and W:

$$TF_{CEBE}(s) = TF_{CEB}(s) - TF_{CEE}(s) \tag{C.15}$$

$$TF_{CECE}(s) = TF_{CEC}(s) - TF_{CEE}(s) \tag{C.16}$$

$$TF_{CECB}(s) = TF_{CEC}(s) - TF_{CEB}(s) \tag{C.17}$$

$$TF_{BEBE}(s) = TF_{BEB}(s) - TF_{BEE}(s) \tag{C.18}$$

$$TF_{BECE}(s) = TF_{BEC}(s) - TF_{BEE}(s) \tag{C.19}$$

$$TF_{BECB}(s) = TF_{BEC}(s) - TF_{BEB}(s) \tag{C.20}$$

$$TF_{CBBE}(s) = TF_{CBB}(s) - TF_{CBE}(s) \tag{C.21}$$

$$TF_{CBCE}(s) = TF_{CBC}(s) - TF_{CBE}(s) \tag{C.22}$$

$$TF_{CBCB}(s) = TF_{CBC}(s) - TF_{CBB}(s) \tag{C.23}$$

Now we can proceed to derive the contributions of IM3 in the collector voltage. Purely third-degree terms were listed in Chapter 4, and here only the cascaded second-degree mechanisms are listed. They are grouped as 21 upconverted envelope terms (V_{CEx}), 21 downconverted harmonic terms (V_{CHx}), and 24 electrothermal terms (V_{CTx}).

The 21 IM3L terms upconverted from the envelope frequency are:

$$V_{CE1}(2\omega_1 - \omega_2) = K_{2GM}^2 \cdot TF_{CEC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CEBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \tag{C.24}$$

$$V_{CE2}(2\omega_1 - \omega_2) = K_{2GO}^2 \cdot TF_{CEC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CECE}(\omega_2 - \omega_1)} \cdot TF(\omega_1)^2 \cdot \overline{TF(\omega_2)} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \tag{C.25}$$

$$V_{CE3}(2\omega_1 - \omega_2) = 1/4 \cdot K_{2GMGO}^2 \cdot TF_{CEC}(2\omega_1 - \omega_2) \cdot [\overline{TF_{CEC}(\omega_2 - \omega_1)} + TF(\omega_1) \cdot \overline{TF_{CEB}(\omega_2 - \omega_1)}] \cdot [TF(\omega_1) + \overline{TF(\omega_2)}] \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \tag{C.26}$$

$$V_{CE4}(2\omega_1 - \omega_2) = K_{2GPI}^2 \cdot TF_{BEC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BEBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \tag{C.27}$$

$$V_{CE5}(2\omega_1 - \omega_2) = j(2\omega_1 - \omega_2) \cdot j(\omega_1 - \omega_2) \cdot K^2{}_{2CPI} \cdot TF_{BEC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BEBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \tag{C.28}$$

$$V_{CE6}(2\omega_1-\omega_2) = j(2\omega_1-\omega_2)\cdot j(\omega_1-\omega_2)\cdot K_{2CBC}^2 \cdot TF_{CBC}(2\omega_1-\omega_2)\cdot \overline{TF_{CBCB}(\omega_2-\omega_1)}\cdot [TF(\omega_1)-1] \cdot [\overline{TF(\omega_2)-1}]\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.29}$$

$$V_{CE7}(2\omega_1-\omega_2) = K_{2GM}\cdot K_{2GO}\cdot TF_{CEC}(2\omega_1-\omega_2) \cdot [\overline{TF_{CEC}(\omega_2-\omega_1)} + TF(\omega_2)\cdot \overline{TF_{CBE}(\omega_2-\omega_1)}]\cdot TF(\omega_1) \cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.30}$$

$$V_{CE8}(2\omega_1-\omega_2) = 1/2\cdot K_{2GM}\cdot K_{2GMGO}\cdot TF_{CEC}(2\omega_1-\omega_2) \cdot \{\overline{TF_{CEC}(\omega_2-\omega_1)} + \overline{TF_{CEB}(\omega_2-\omega_1)} \cdot [2\cdot TF(\omega_1)+\overline{TF(\omega_2)}]\}\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.31}$$

$$V_{CE9}(2\omega_1-\omega_2) = K_{2GM}\cdot K_{2GPI}\cdot [TF_{BEC}(2\omega_1-\omega_2) \cdot \overline{TF_{CEBE}(\omega_2-\omega_1)} + TF_{CEC}(2\omega_1-\omega_2)\cdot \overline{TF_{BEBE}(\omega_2-\omega_1)}] \cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.32}$$

$$V_{CE10}(2\omega_1-\omega_2) = K_{2GM}\cdot K_{2CPI}\cdot [j(2\omega_1-\omega_2) \cdot TF_{BEC}(2\omega_1-\omega_2)\cdot \overline{TF_{CEB}(\omega_2-\omega_1)} + j(\omega_1-\omega_2) \cdot TF_{CEC}(2\omega_1-\omega_2)\cdot \overline{TF_{BEB}(\omega_2-\omega_1)}]\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.33}$$

$$V_{CE11}(2\omega_1-\omega_2) = K_{2GM}\cdot K_{2CBC}\cdot [TF(\omega_1)-1] \cdot [j(2\omega_1-\omega_2)\cdot TF_{CBC}(2\omega_1-\omega_2)\cdot \overline{TF_{CECB}(\omega_2-\omega_1)} + j(\omega_1-\omega_2)\cdot TF_{CEC}(2\omega_1-\omega_2)\cdot \overline{TF_{CBBE}(\omega_2-\omega_1)} \cdot (\overline{TF(\omega_2)}-1)]\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.34}$$

$$V_{CE12}(2\omega_1-\omega_2) = 1/2\cdot K_{2GO}\cdot K_{2GMGO}\cdot TF_{CEC}(2\omega_1-\omega_2) \cdot \{TF(\omega_1)\cdot \overline{TF(\omega_2)}\cdot \overline{TF_{CBE}(\omega_2-\omega_1)} + \overline{TF_{CEC}(\omega_2-\omega_1)} \cdot [2\cdot \overline{TF(\omega_2)} + TF(\omega_1)]\}\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.35}$$

$$V_{CE13}(2\omega_1-\omega_2) = K_{2GO}\cdot K_{2GPI}\cdot\{TF_{BEC}(2\omega_1-\omega_2) \cdot \overline{TF(\omega_2)}\cdot \overline{TF_{CEB}(\omega_2-\omega_1)} + TF_{CEC}(2\omega_1-\omega_2) \cdot \overline{TF_{BCE}(\omega_2-\omega_1)}\}\cdot TF(\omega_1)\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.36}$$

$$V_{CE14}(2\omega_1-\omega_2) = K_{2GO}\cdot K_{2CPI}\cdot\{j(2\omega_1-\omega_2) \cdot TF_{BC}(2\omega_1-\omega_2)\cdot \overline{TF(\omega_2)}\cdot \overline{TF_{CBE}(\omega_2-\omega_1)} + j(\omega_1-\omega_2) \cdot TF_{CC}(2\omega_1-\omega_2)\cdot \overline{TF_{BCE}(\omega_2-\omega_1)}\}\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.37}$$

$$V_{CE15}(2\omega_1-\omega_2) = K_{2GO}\cdot K_{2CBC}\cdot TF(\omega_1)\cdot[TF(\omega_1)-1] \cdot[j(2\omega_1-\omega_2)\cdot TF_{CBC}(2\omega_1-\omega_2)\cdot \overline{TF(\omega_2)} \cdot \overline{TF_{CECB}(\omega_2-\omega_1)} + j(\omega_1-\omega_2)\cdot TF_{CEC}(2\omega_1-\omega_2) \cdot \overline{TF_{CBCE}(\omega_2-\omega_1)}\cdot(\overline{TF(\omega_2)}-1)]\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.38}$$

$$V_{CE16}(2\omega_1-\omega_2) = 1/2\cdot K_{2GPI}\cdot K_{2GMGO}\cdot\{TF_{CC}(2\omega_1-\omega_2) \cdot[\overline{TF_{BBE}(\omega_2-\omega_1)}\cdot TF(\omega_1)+\overline{TF_{BCE}(\omega_2-\omega_1)}] + TF_{BC}(2\omega_1-\omega_2)\cdot \overline{TF_{CBE}(\omega_2-\omega_1)}\cdot[TF(\omega_1)+\overline{TF(\omega_2)}]\} \cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.39}$$

$$V_{CE17}(2\omega_1-\omega_2) = K_{2GPI}\cdot K_{2CBC}\cdot[TF(\omega_1)-1] \cdot[j(2\omega_1-\omega_2)\cdot TF_{CBC}(2\omega_1-\omega_2)\cdot \overline{TF_{BECB}(\omega_2-\omega_1)} + j(\omega_1-\omega_2)\cdot TF_{BEC}(2\omega_1-\omega_2)\cdot \overline{TF_{CBBE}(\omega_2-\omega_1)} \cdot(\overline{TF(\omega_2)}-1)]\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.40}$$

$$V_{CE18}(2\omega_1-\omega_2) = 1/2\cdot K_{2CPI}\cdot K_{2GMGO}\cdot\{j(\omega_1-\omega_2) \cdot TF_{CC}(2\omega_1-\omega_2)\cdot[\overline{TF_{BBE}(\omega_2-\omega_1)}\cdot TF(\omega_1) + \overline{TF_{BCE}(\omega_2-\omega_1)}] + j(2\omega_1-\omega_2)\cdot TF_{BC}(2\omega_1-\omega_2) \cdot \overline{TF_{CBE}(\omega_2-\omega_1)}\cdot[TF(\omega_1)+\overline{TF(\omega_2)}]\}\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.41}$$

$$V_{CE19}(2\omega_1 - \omega_2) = j(2\omega_1 - \omega_2) \cdot j(\omega_1 - \omega_2) \cdot K_{2CPI} \cdot K_{2CBC} \cdot [TF(\omega_1) - 1] \cdot [TF_{CBC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BECB}(\omega_2 - \omega_1)} + TF_{BEC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CBBE}(\omega_2 - \omega_1)} \cdot (\overline{TF(\omega_2)} - 1)] \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.42)}$$

$$V_{CE20}(2\omega_1 - \omega_2) = j(3\omega_1 - 2\omega_2) \cdot K_{2GPI} \cdot K_{2CPI} \cdot TF_{BC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.43)}$$

$$V_{CE21}(2\omega_1 - \omega_2) = 1/2 \cdot K_{2GMGO} \cdot K_{2CBC} \cdot [TF(\omega_1) - 1] \cdot [j(2\omega_1 - \omega_2) \cdot TF_{CBC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CECB}(\omega_2 - \omega_1)} \cdot (TF(\omega_1) + \overline{TF(\omega_2)}) + j(\omega_1 - \omega_2) \cdot TF_{CEC}(2\omega_1 - \omega_2) \cdot ([\overline{TF_{CBBE}(\omega_2 - \omega_1)} \cdot TF(\omega_1) + TF_{CBCE}(\omega_2 - \omega_1)] \cdot [\overline{TF(\omega_2)} - 1]] \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.44)}$$

The 21 IM3L terms downconverted from the second harmonic frequency are

$$V_{CH1}(2\omega_1 - \omega_2) = 1/2 \cdot K_{2GM}^2 \cdot TF_{CEC}(2\omega_1 - \omega_2) \cdot TF_{CEBE}(2\omega_1) \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.45)}$$

$$V_{CH2}(2\omega_1 - \omega_2) = 1/2 \cdot K_{2GO}^2 \cdot TF_{CEC}(2\omega_1 - \omega_2) \cdot TF_{CECE}(2\omega_1) \cdot TF(\omega_1)^2 \cdot \overline{TF(\omega_2)} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.46)}$$

$$V_{CH3}(2\omega_1 - \omega_2) = 1/4 \cdot K_{2GMGO}^2 \cdot TF_{CC}(2\omega_1 - \omega_2) \cdot [TF_{CCE}(2\omega_1) + \overline{TF(\omega_2)} \cdot TF_{CBE}(2\omega_1)] \cdot TF(\omega_1) \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.47)}$$

$$V_{CH4}(2\omega_1 - \omega_2) = 1/2 \cdot K_{2GPI}^2 \cdot TF_{BEC}(2\omega_1 - \omega_2) \cdot TF_{BEBE}(2\omega_1) \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.48)}$$

$$V_{CH5}(2\omega_1-\omega_2) = j(2\omega_1-\omega_2)\cdot j\omega_1\cdot K_{2CPI}\,\hat{}\,2 \cdot TF_{BEC}(2\omega_1-\omega_2)\cdot TF_{BEBE}(2\omega_1)\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.49}$$

$$V_{CH6}(2\omega_1-\omega_2) = j(2\omega_1-\omega_2)\cdot j\omega_1\cdot K_{2CBC} \,\hat{}\, 2\cdot TF_{CBC}(2\omega_1-\omega_2)\cdot TF_{CBCB}(2\omega_1)\cdot [TF(\omega_1)-1]\,\hat{}\,2 \cdot [\overline{TF(\omega_2)-1}]\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.50}$$

$$V_{CH7}(2\omega_1-\omega_2)\cdot 1/2\cdot K_{2GM}\cdot K_{2GO}\cdot TF_{CC}(2\omega_1-\omega_2) \cdot [\overline{TF(\omega_2)}\cdot TF_{CCE}(2\omega_1)+TF(\omega_1)^2\cdot TF_{CBE}(2\omega_1)] \cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.51}$$

$$V_{CH8}(2\omega_1-\omega_2) = 1/4\cdot K_{2GM}\cdot K_{2GMGO}\cdot TF_{CC}(2\omega_1-\omega_2) \cdot \{TF_{CCE}(2\omega_1)+TF_{CBE}(2\omega_1)\cdot [2\cdot TF(\omega_1)+\overline{TF(\omega_2)}]\} \cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.52}$$

$$V_{CH9}(2\omega_1-\omega_2) = 1/2\cdot K_{2GM}\cdot K_{2GPI}\cdot [TF_{BC}(2\omega_1-\omega_2) \cdot TF_{CBE}(2\omega_1)+TF_{CC}(2\omega_1-\omega_2)\cdot F_{BBE}(2\omega_1)] \cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.53}$$

$$V_{CH10}(2\omega_1-\omega_2) = K_{2GM}\cdot K_{2CPI}\cdot [j(2\omega_1-\omega_2) \cdot TF_{BC}(2\omega_1-\omega_2)\cdot TF_{CBE}(2\omega_1)+j\omega_1\cdot TF_{CC}(2\omega_1-\omega_2) \cdot TF_{BBE}(2\omega_1)]\cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.54}$$

$$V_{CH11}(2\omega_1-\omega_2) = K_{2GM}\cdot K_{2CBC}\cdot [1/2\cdot j(2\omega_1-\omega_2) \cdot TF_{CBC}(2\omega_1-\omega_2)\cdot TF_{CECB}((2\omega_1)\cdot [1-\overline{TF(\omega_2)}]) + j\omega_1\cdot TF_{CEC}(2\omega_1-\omega_2)\cdot TF_{CBBE}(2\omega_1)\cdot [TF(\omega_1)-1]^2] \cdot V_{BE}(\omega_1)^2\cdot \overline{V_{BE}(\omega_2)} \tag{C.55}$$

$$V_{CH12}(2\omega_1-\omega_2) = 1/4 \cdot K_{2GO} \cdot K_{2GMGO} \cdot TF_{CC}(2\omega_1-\omega_2) \quad \text{(C.56)}$$
$$\cdot \{TF(\omega_1) \cdot \overline{TF(\omega_2)} \cdot TF_{CBE}(2\omega_1)$$
$$+ TF_{CCE}(2\omega_1) \cdot [2 \cdot \overline{TF(\omega_2)} + TF(\omega_1)]\} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)}$$

$$V_{CH13}(2\omega_1-\omega_2) = 1/2 \cdot K_{2GO} \cdot K_{2GPI} \quad \text{(C.57)}$$
$$\cdot \{TF_{BC}(2\omega_1-\omega_2) \cdot TF(\omega_1)^2 \cdot TF_{CBE}(2\omega_1) + TF_{CC}(2\omega_1-\omega_2)$$
$$\cdot TF_{BCE}(2\omega_2) \cdot \overline{TF(\omega_2)}\} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)}$$

$$V_{CH14}(2\omega_1-\omega_2) = 1/2 \cdot K_{2GO} \cdot K_{2CPI} \cdot \{j(2\omega_1-\omega_2) \quad \text{(C.58)}$$
$$\cdot TF_{BC}(2\omega_1-\omega_2) \cdot TF(\omega_1)^2 \cdot TF_{CBE}(2\omega_1)$$
$$+ j\omega_1 \cdot TF_{CC}(2\omega_1-\omega_2) \cdot TF_{BCE}(2\omega_1)\} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)}$$

$$V_{CH15}(2\omega_1-\omega_2) = 1/2 \cdot K_{2GO} \cdot K_{2CBC} \cdot [j(2\omega_1-\omega_2) \quad \text{(C.59)}$$
$$\cdot TF_{CBC}(2\omega_1-\omega_2) \cdot TF(\omega_1)^2 \cdot TF_{CECB}(2\omega_1)$$
$$\cdot TF(\omega_1)^2 \cdot [\overline{TF(\omega_2)} - 1] + j2\omega_1 \cdot TF_{CEC}(2\omega_1-\omega_2)$$
$$\cdot TF_{CBCE}(2\omega_1) \cdot \overline{TF(\omega_2)} \cdot [TF(\omega_1) - 1]^2] \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)}$$

$$V_{CH16}(2\omega_1-\omega_2) = 1/4 \cdot K_{2GPI} \cdot K_{2GMGO} \quad \text{(C.60)}$$
$$\cdot \{TF_{CC}(2\omega_1-\omega_2) \cdot [TF_{BBE}(2\omega_1) \cdot \overline{TF(\omega_2)} + TF_{BCE}(2\omega_1)]$$
$$+ TF_{BC}(2\omega_1-\omega_2) \cdot TF_{CBE}(2\omega_1) \cdot TF(\omega_1)\} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)}$$

$$V_{CH17}(2\omega_1-\omega_2) = 1/2 \cdot K_{2GPI} \cdot K_{2CBC} \cdot \{j(2\omega_1-\omega_2) \quad \text{(C.61)}$$
$$\cdot TF_{CBC}(2\omega_1-\omega_2) \cdot TF(\omega_1)^2 \cdot TF_{BECB}((2\omega_1) \cdot [\overline{TF(\omega_2)} - 1])$$
$$+ j2\omega_1 \cdot TF_{BEC}(2\omega_1-\omega_2) \cdot TF_{CBBE}(2\omega_1) \cdot [TF(\omega_1) - 1]^2\}$$
$$\cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)}$$

$$V_{CH18}(2\omega_1-\omega_2) = 1/2 \cdot K_{2CPI} \cdot K_{2GMGO} \cdot \{j\omega_1 \cdot TF_{CC}(2\omega_1-\omega_2) \cdot [TF_{BBE}(2\omega_1) \cdot \overline{TF(\omega_2)} + TF_{BCE}(2\omega_1)] + j(2\omega_1-\omega_2) \cdot TF_{BC}(2\omega_1-\omega_2) \cdot TF_{CBE}(2\omega_1) \cdot TF(\omega_1)\} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.62)}$$

$$V_{CH19}(2\omega_1-\omega_2) = j(2\omega_1-\omega_2) \cdot K_{2CPI} \cdot K_{2CBC} \cdot \{j\omega_1 \cdot TF_{CBC}(2\omega_1-\omega_2) \cdot TF_{BECB}(2\omega_1) \cdot [\overline{TF(\omega_2)} - 1] + j\omega_1 \cdot TF_{BEC}(2\omega_1-\omega_2) \cdot TF_{CBBE}(2\omega_1) \cdot [TF(\omega_1)-1]^2\} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.63)}$$

$$V_{CH20}(2\omega_1-\omega_2) = 1/2 \cdot j(4\omega_1-\omega_2) \cdot K_{2GPI} \cdot K_{2CPI} \cdot TF_{BC}(2\omega_1-\omega_2) \cdot TF_{BBE}(2\omega_1) \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.64)}$$

$$V_{CH21}(2\omega_1-\omega_2) = 1/2 \cdot K_{2CBC} \cdot K_{2GMGO} \cdot \{j\omega_1 \cdot TF_{CEC}(2\omega_1-\omega_2) \cdot [TF_{CBBE}(2\omega_1) \cdot \overline{TF(\omega_2)} + TF_{CBCE}(2\omega_1)] \cdot [TF(\omega_1)-1]^2 + j(2\omega_1-\omega_2) \cdot TF_{CBC}(2\omega_1-\omega_2) \cdot TF_{CECB}(2\omega_1) \cdot TF(\omega_1) \cdot [\overline{TF(\omega_2)} - 1]\} \cdot V_{BE}(\omega_1)^2 \cdot \overline{V_{BE}(\omega_2)} \quad \text{(C.65)}$$

Finally, the 24 electrothermal second-degree terms can be expressed with the help of envelope frequency temperature $T_x = T_x(\omega_2–\omega_1)$. Here a subscript G refers to the temperature of the $g_m/g_o/g_{pi}$, C to C_{pi}, CBC to C_{bc}

$$K_{2GM} \cdot K_{2GPT} \cdot TF_{CC}(2\omega_1-\omega_2) \cdot \overline{TF_{BEBE}(\omega_2-\omega_1)} \cdot V_{BE}(\omega_1) \cdot T_G \quad \text{(C.66)}$$

$$j(\omega_1-\omega_2) \cdot K_{2GM} \cdot K_{2CPT} \cdot TF_{CEC}(2\omega_1-\omega_2) \cdot \overline{TF_{BEBE}(\omega_2-\omega_1)} \cdot V_{BE}(\omega_1) \cdot T_C \quad \text{(C.67)}$$

$$K_{2GM} \cdot K_{2GMT} \cdot TF_{CC}(2\omega_1-\omega_2) \cdot \overline{TF_{CEBE}(\omega_2-\omega_1)} \cdot V_{BE}(\omega_1) \cdot T_M \quad \text{(C.68)}$$

$$j(\omega_1-\omega_2)\cdot K_{2GM}\cdot K_{2CBCT}\cdot TF_{CEC}(2\omega_1-\omega_2)\cdot\overline{TF_{CBBE}(\omega_2-\omega_1)}\cdot V_{BE}(\omega_1)\cdot T_{CBC} \tag{C.69}$$

$$K_{2GO}\cdot K_{2GPT}\cdot TF_{CEC}(2\omega_1-\omega_2)\cdot TF(\omega_1)\cdot\overline{TF_{BECE}(\omega_2-\omega_1)}\cdot V_{BE}(\omega_1)\cdot T_G \tag{C.70}$$

$$j(\omega_1-\omega_2)\cdot K_{2GO}\cdot K_{2CPT}\cdot TF_{CEC}(2\omega_1-\omega_2)\cdot TF(\omega_1)\cdot\overline{TF_{BBE}(\omega_2-\omega_1)}\cdot V_{BE}(\omega_1)\cdot T_C \tag{C.71}$$

$$K_{2GO}\cdot K_{2GMT}\cdot TF_{CEC}(2\omega_1-\omega_2)\cdot TF(\omega_1)\cdot\overline{TF_{CECE}(\omega_2-\omega_1)}\cdot V_{BE}(\omega_1)\cdot T_M \tag{C.72}$$

$$j(\omega_1-\omega_2)\cdot K_{2GO}\cdot K_{2CBCT}\cdot TF_{CEC}(2\omega_1-\omega_2)\cdot TF(\omega_1)\cdot\overline{TF_{CBCE}(\omega_2-\omega_1)}\cdot V_{BE}(\omega_1)\cdot T_{CBC} \tag{C.73}$$

$$1/2\cdot K_{2GMGO}\cdot K_{2GPT}\cdot TF_{CC}(2\omega_1-\omega_2)\cdot TF(\omega_1)\cdot[TF(\omega_1)\cdot\overline{TF_{BBE}(\omega_2-\omega_1)}+\overline{TF_{BCE}(\omega_2-\omega_1)}]\cdot V_{BE}(\omega_1)\cdot T_G \tag{C.74}$$

$$1/2\cdot j(\omega_1-\omega_2)\cdot K_{2GMGO}\cdot K_{2CPT}\cdot TF_{CC}(2\omega_1-\omega_2)\cdot TF(\omega_1)\cdot[TF(\omega_1)\cdot\overline{TF_{BBE}(\omega_2-\omega_1)}+\overline{TF_{BCE}(\omega_2-\omega_1)}]\cdot V_{BE}(\omega_1)\cdot T_C \tag{C.75}$$

$$1/2\cdot K_{2GMGO}\cdot K_{2GMT}\cdot TF_{CC}(2\omega_1-\omega_2)\cdot TF(\omega_1)\cdot[TF(\omega_1)\cdot\overline{TF_{BBE}(\omega_2-\omega_1)}+\overline{TF_{BCE}(\omega_2-\omega_1)}]\cdot V_{BE}(\omega_1)\cdot T_M \tag{C.76}$$

$$1/2\cdot j(\omega_1-\omega_2)\cdot K_{2GMGO}\cdot K_{2CBCT}\cdot TF_{CEC}(2\omega_1-\omega_2)\cdot TF(\omega_1)\cdot[TF(\omega_1)\cdot\overline{TF_{CBBE}(\omega_2-\omega_1)}+\overline{TF_{CBCE}(\omega_2-\omega_1)}]\cdot V_{BE}(\omega_1)\cdot T_{CBC} \tag{C.77}$$

$$K_{2GPI} \cdot K_{2GPT} \cdot TF_{BC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1) \cdot T_G \tag{C.78}$$

$$j(\omega_1 - \omega_2) \cdot K_{2GPI} \cdot K_{2CPT} \cdot TF_{BC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1) \cdot T_C \tag{C.79}$$

$$K_{2GPI} \cdot K_{2GMT} \cdot TF_{BC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1) \cdot T_M \tag{C.80}$$

$$j(\omega_1 - \omega_2) \cdot K_{2GPI} \cdot K_{2CBCT} \cdot TF_{BEC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CBBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1) \cdot T_{CBC} \tag{C.81}$$

$$K_{2CPI} \cdot K_{2GPT} \cdot TF_{BC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1) \cdot T_G \tag{C.82}$$

$$j(2\omega_1 - \omega_2) \cdot j(\omega_1 - \omega_2) \cdot K_{2CPI} \cdot K_{2CPT} \cdot TF_{BC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1) \cdot T_C \tag{C.83}$$

$$K_{2CPI} \cdot K_{2GMT} \cdot TF_{BC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1) \cdot T_M \tag{C.84}$$

$$j(2\omega_1 - \omega_2) \cdot j(\omega_1 - \omega_2) \cdot K_{2CPI} \cdot K_{2CBCT} \cdot TF_{BC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CBBE}(\omega_2 - \omega_1)} \cdot V_{BE}(\omega_1) \cdot T_{CBC} \tag{C.85}$$

$$j(2\omega_1 - \omega_2) \cdot K_{2CBC} \cdot K_{2GMT} \cdot TF_{CBC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CECB}(\omega_2 - \omega_1)} \cdot [TF(\omega_1) - 1] \cdot V_{BE}(\omega_1) \cdot T_M \tag{C.86}$$

$$j(2\omega_1 - \omega_2) \cdot K_{2CBC} \cdot K_{2GPT} \cdot TF_{CBC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BECB}(\omega_2 - \omega_1)} \cdot [TF(\omega_1) - 1] \cdot V_{BE}(\omega_1) \cdot T_{GPI} \tag{C.87}$$

$$j(2\omega_1 - \omega_2) \cdot j(\omega_1 - \omega_2) \cdot K_{2CBC} \cdot K_{2CPIT} \cdot TF_{CBC}(2\omega_1 - \omega_2) \cdot \overline{TF_{BECB}(\omega_2 - \omega_1)} \cdot [TF(\omega_1) - 1] \cdot V_{BE}(\omega_1) \cdot T_{CPI} \quad \text{(C.88)}$$

$$j(2\omega_1 - \omega_2) \cdot j(\omega_1 - \omega_2) \cdot K_{2CBC} \cdot K_{2CBCT} \cdot TF_{CBC}(2\omega_1 - \omega_2) \cdot \overline{TF_{CBCB}(\omega_2 - \omega_1)} \cdot [TF(\omega_1) - 1] \cdot V_{BE}(\omega_1) \cdot T_{CBC} \quad \text{(C.89)}$$

Appendix D: About the Measurement Setups

Various test setups and circuit techniques are presented in this book. While the core text is more focused on explaining the ideas of the circuit techniques and the results of the measurements, this appendix explains the techniques in more detail to give the reader some hints on how to build similar setups.

All the measurement setups and predistorter devices are built using commercial components, and a few words can first be said here about the building blocks and instruments. At the heart of the test setups there are several RF signal generators, all locked to the same frequency reference. Usually three generators are used, two of them generating the fundamental two-tone test signal, while one generates an injection signal at an intermodulation frequency, for example, an envelope (f_2–f_1) or IM3 ($2f_2$–f_1 or $2f_1$–f_2), depending on the type of the test setup. Locking to the same frequency reference is necessary to avoid huge phase drifting, but still a slow drift was seen, and it was necessary to calibrate the phase regularly. Another serious problem is related to the output power control, as both the continuous amplitude control and the step attenuators making larger changes affect not only the amplitude, but also the phase of the output signal phase. Hence, it is necessary to calibrate the phase vs. amplitude dependency of the signal generator. A third well-known problem is the generation of intermodulation tones in the signal generators when the outputs of several generators are combined. The attenuation of the combiners reduces the distortion a little, and if this is not sufficient, circulators are needed to avoid the direct coupling between the signal generators.

All the power combiners/splitters and mixers are Mini-Circuits components [1], and care is needed to avoid unwanted spectral components and to cope with the limited linear range and isolation of the mixers when building the test circuits. In many test setups, for example, the RF signal is squared down to the envelope frequency and spectral components around

the second harmonic have to be filtered. Isolation of the mixers is extremely important in cases where two signals that have significantly different power levels are mixed. Many problems in power sweeps also arise from the amplitude of the mixing products, because these are not always proportional to the product of the two inputs. Thus, the amplitude levels must be chosen carefully.

The most often used building block in this book is the polynomial RF predistorter, shown in Figure D.1. The input signal, being a two-tone signal in all cases here, is split into four branches. The uppermost branch is for the linear signal (a_1) and it goes directly to the output power combiner of the predistorter. The last three branches are mixed together to produce first the envelope signal and then the third-order signal, the amplitude and phase of which is tuned to find the sufficient value of a_3 for maximum cancellation. The second-order envelope signal is filtered to remove the second harmonics, so just the envelope signal is mixed back to the fundamental band. In some cases, for example when using envelope filtering, the envelope is also slightly filtered to produce sufficient memory effects inside the predistorter, and when using the envelope injection technique, the envelope signal is taken out to be fed to the input of the amplifier separately.

The purpose of the very simplified polynomial predistorter is just to produce the IM3 sidebands that have a controllable amplitude and phase. No special attention has been paid to making sure that the fundamental gain

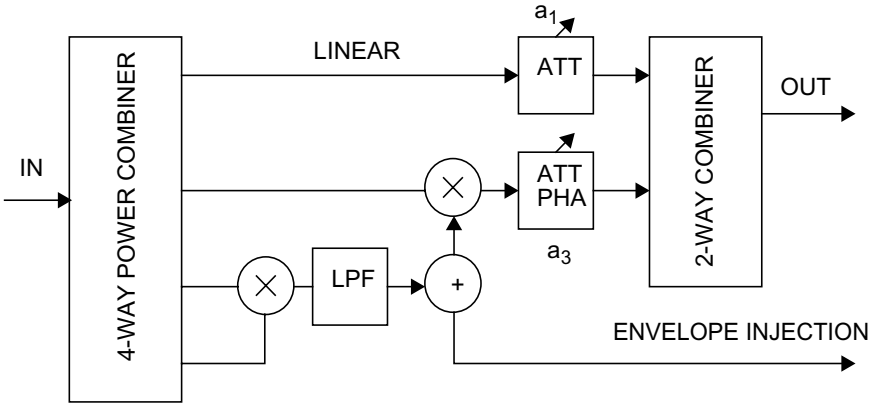


Figure D.1 The structure of a discrete polynomial RF predistorter circuit with envelope injection output.

expansion of the predistorter is correct to cancel the compression of the amplifier.

In theory, if both the predistorter and the amplifier are memoryless (i.e., can be modeled as polynomials), the tuning of a_3 of the predistorter guarantees simultaneous cancellation of IM3 and correction of the fundamental signal. In practice this is not true, and in many cases different coefficients are used in polynomial predistorters to correct the fundamental signal and IM3. For example, if only the second-order signal of the predistorter is filtered to remove the second harmonics, the predistorter is no more memoryless, and simultaneous correction for both fundamental signal and IM3 cannot be obtained, and even the circuit does not otherwise show memory at all. However, since the purpose of this book is to study the memory effects of IM3 components, no special attention for fundamental signals is given so that the test circuitry is as simple as possible.

Reference

[1] http://www.minicircuits.com/.

Glossary

a_n	nth degree nonlinearity coefficient.
ac	Alternating current.
ACPR	Adjacent channel power ratio. The amount of power leaking to the next channel compared to the power of the own channel. Measured using modulated data and real raised cosine filters in the receiver.
AM-AM	Amplitude-dependent amplitude distortion.
AM-PM	Amplitude-dependent phase distortion. Both AM-AM and AM-PM are normally measured as single-tone power sweep measurements.
AMPS	Advanced mobile phone system. U.S. first generation mobile phone standard employing analog FM modulation.
BJT	Bipolar junction transistor.
BSIM	Berkeley short-channel IGFET model.
CDMA	Code division multiple access. Different users transmit at the same frequency and the same time but are separated by orthogonal spreading codes.
CE	Common emitter. A BJT amplifier where the emitter is grounded.
CF	Crest factor. Ratio between the peak and average powers. To avoid clipping of the peak powers, high CF requires high backoff.
CS	Common source. An FET amplifier where the source is grounded.
CW	Continuous wave. Nonpulsed sinusoid signal.
dBc	Power in decibels compared to the carrier or fundamental signal.
dc	Direct current.

Degree	Of nonlinearity. In v^N, N is the degree of the nonlinearity. Defines the shape of the nonlinear characteristics.
DSP	Digital signal processing.
DUT	Device under test.
EER	Envelope elimination and restoration. A linearization technique based on the use of constant-envelope amplifier and modulated power supply.
Envelope	The amplitude variation of the carrier. In this book, envelope mostly refers to the second-order rectification result, that in a two-tone test appears at the difference frequency f_2–f_1. Also called the video or beat frequency.
FET	Field effect transistor.
FM	Frequency modulation.
GSM	Global System for Mobile communications. Originally European second generation mobile phone standard using constant-envelope GMSK modulation and 1/8 duty cycle TDMA.
GMSK	Gaussian minimum shift keying. A constant-envelope modulation scheme with 1.3 bit/s/Hz spectral efficiency.
HB	Harmonic balance, a steady-state simulation algorithm.
HBT	Heterojunction bipolar transistor. Built using different bandgap materials in the base and emitter regions.
HD	Harmonic distortion, appears at the multiples of the input frequencies.
IC	Integrated circuit.
IF	Intermediate frequency.
I-V	Current-voltage characteristic.
IGFET	Insulated gate FET.
IM	Intermodulation distortion. In a two-tone test, appears at frequencies Kf_1+Lf_2, where K and L are nonzero integers.
IM3	Third-order intermodulation.
IM3L, IM3H	Lower and higher IM3 tones at $2f_1$–f_2 and $2f_2$–f_1 (f_2>f_1).
IM5	Fifth-order intermodulation.
K_n	nth-degree nonlinearity coefficient.
LDMOS	Laterally diffused field effect transistor.
LMSE	Least mean square error. Minimizing the rms error.
LNA	Low noise amplifier.
Load-pull	Search of optimum performance by varying either the fundamental or harmonic load impedance.

Memory effect	IM distortion is not constant but its phase or amplitude varies with the distance to the center of the channel.
MESFET	Metal-semiconductor field effect transistor.
MET	Motorola electro thermal model.
MOSFET	Metal-oxide-semiconductor field effect transistor.
MNA	Modified nodal analysis. Commonly used technique in simulators, where most of the circuit is described by current equations in the nodes, but some branch currents are added as variables to model voltage sources and inductors, for example.
Modulation frequency	Varying rate of the envelope of the carrier. In a two-tone test the same as the tone spacing.
NMT	Nordic Mobile Telephone. Scandinavian first generation mobile phone standard using analog FM modulation.
NWA	Network analyzer.
Order	Of distortion product. Defines how many fundamental tones need to be multiplied to create an *N*th-order distortion tone. The amplitude of *N*th-order distortion is proportional to A_{in}^{N}, where A_{in} is the input amplitude.
P_{1dB}	1-dB compression point. Power level when the large signal gain has dropped by 1 dB.
P_{IIP3}	Input intercept point. Extrapolated input level where the fundamental and IM3 amplitudes are equal.
PA	Power amplifier.
PAE	Power added efficiency.
RF	Radio frequency.
QAM	Quadrature amplitude modulation. In QAM-N modulation I and Q amplitudes are modulated so that altogether, N different constellation points are generated. High spectral efficiency (ideally $\log_2$(N) bit/s/Hz) but also high crest factor.
QPSK	Quadrature phase shift keying. A varying envelope modulation scheme achieving ideally 2 bit/s/Hz spectral efficiency.
Q-V	Charge-voltage characteristic.
Source pull	Search of optimum performance by varying the (here mostly baseband) driving impedance.
TDMA	Time division multiple access. Different users transmit at the same frequency but in different time slots. As the transmitter (of the terminal) can be off most of the time, the overall efficiency is improved.

TF	Transfer function.
TPF	Thermal power feedback. Instantaneous power dissipation varies the junction temperature and hence the gain of the amplifier, causing IM3 distortion.
Transimpedance	Distortion is modeled here as excess current sources, and to get the distortion voltages at certain nodes the currents need to be multiplied by transimpedance transfer functions. Note that even if the distortion current is small but the transimpedance gain is high, it may still cause a high amount of distortion.
TRL	Through-reflect-line calibration method.
VCCS	Voltage-controlled current source. Transconductance element.
VIOMAP	Volterra input output map.
WCDMA	Wideband code division multiple access. International third generation mobile communicator standard, where users share the same channel and are separated with orthogonal spreading codes. This results in a high crest factor, QAM-like modulation in the base station transmitter.
Z_{TH}	Thermal impedance.

About the Authors

Joel Vuolevi received the diploma engineer and doctor of technology degrees in electrical engineering from the University of Oulu, Oulu, Finland, in 1998 and 2001, respectively. From 1997 to 1998, he was an RF design engineer with Nokia Mobile Phones. In 1998, he joined the Electronics Laboratory at the University of Oulu, where he worked as a postgraduate student, an acting professor, and a postdoctoral researcher. His research interests lie in the field of analysis, measurement, and cancellation of distortion, and especially memory effects in RF power amplifiers. He has authored or coauthored numerous published papers on these topics. In 2002 he joined RF Integrated Corporation in Irvine, California. His current technical interests are in the design of linear power amplifiers for future telecommunications systems.

Timo Rahkonen received the diploma engineer, licentiate, and doctor of technology degrees from the University of Oulu, Oulu, Finland, in 1986, 1991, and 1994, respectively, all related to the design of integrated circuits for measuring short time intervals. He is currently a professor of circuit theory and circuit design at the University of Oulu, where he conducts research on nonlinear analysis, linearization of RF power amplifiers, and error-correction techniques for A/D and D/A converters. He has been a member of IEEE since 1988 and has authored or coauthored more than 100 published papers.

Index